Séminaire - Montpellier
du 2 au 4 Septembre 1992

INITIATION

AUX

TECHNIQUES GRAPHIQUES

Mission Biométrie
CIRAD

Séminaire - Montpellier
du 2 au 4 Septembre 1992

INITIATION

AUX

TECHNIQUES GRAPHIQUES

« Le problème avec Möbius, c'est qu'il ne voit qu'un seul côté de la question. »

Mission Biométrie
CIRAD

Sommaire

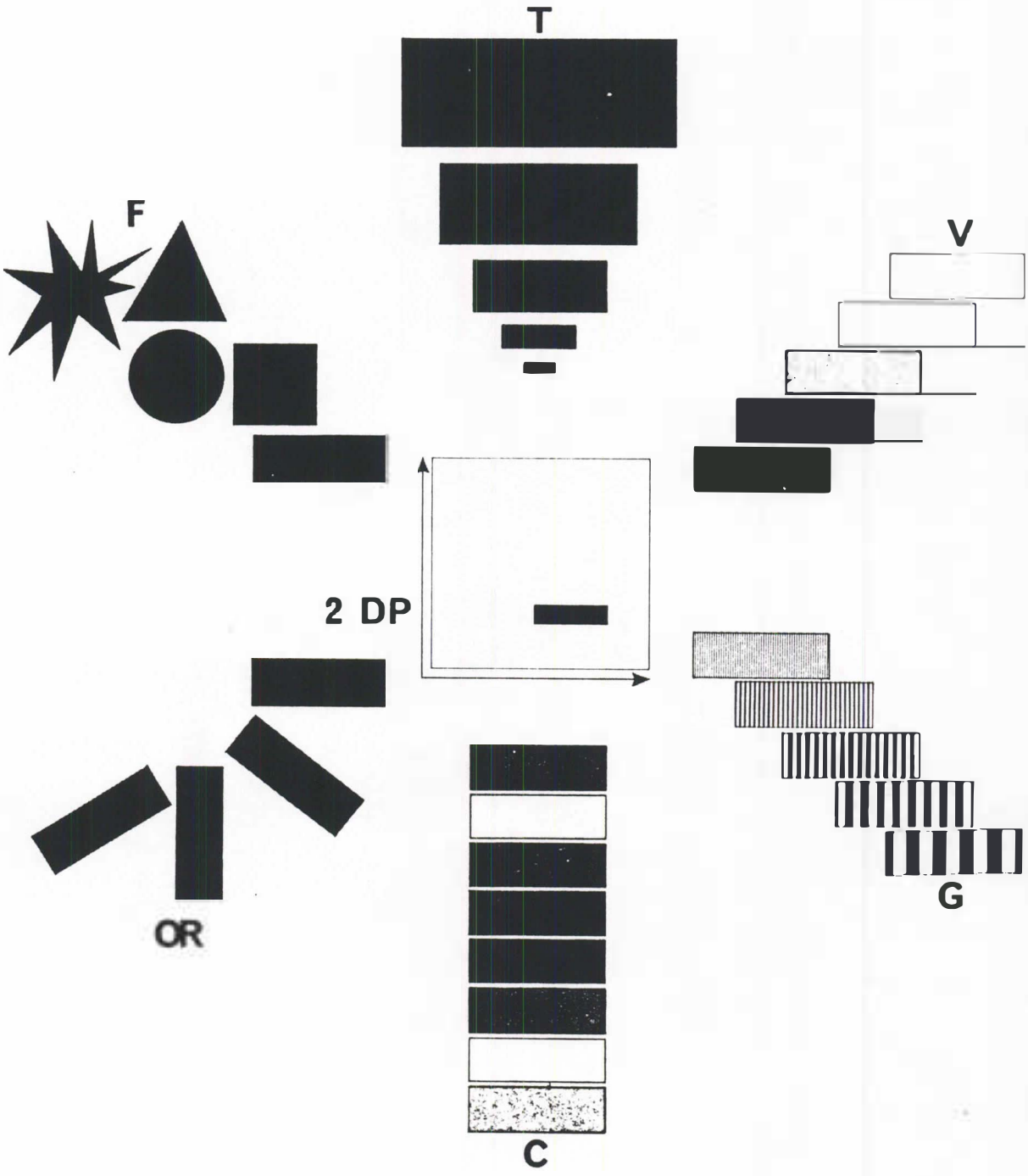
* * *

LES MOYENS DU SYSTEME GRAPHIQUE

Septembre 1992

Jean-Claude BERGONZINI

BIOMETRIE
CIRAD - Forêt


T
F
V
2 DP
OR
G
C

BIOM/JCB/AG

LES MOYENS DU SYSTEME GRAPHIQUE

De quoi disposons-nous ? ... de taches positionnées dans un espace à deux dimensions... le plan.

Ces taches peuvent faire l'objet de huit variations visuelles :

la taille	la valeur	le grain	la couleur	l'orientation	la forme

A - LE PLAN

1. L'implantation

On appelle implantation les trois constituants élémentaires de la géométrie.

le point : Il n'a pas de dimension mais une position. La tache qui le représente peut utiliser les huit variations visuelles évoquées ci-dessus.

la ligne : Elle n'a pas de surface mais une longueur et une position. La tache qui la représente ne peut varier suivant les huit variations visuelles évoquées ci-dessus avec des restrictions sur l'orientation et la forme.

la zone ou surface : La tache qui la représente ne peut pas varier d'orientation ou de forme.

.../...

L'ANALYSE DES QUANTITÉS A REPRÉSENTER, première conséquence de l'implantation.

Lorsque les classes sont de dimensions variables, la représentation des quantités affectées à ces classes doit tenir compte : 1) de l'implantation ponctuelle, linéaire ou zonale des classes ; 2) de la nature Q ou QS des quantités à représenter (p. 38).
Soit l'information suivante concernant quatre communes (classes) A B C D :

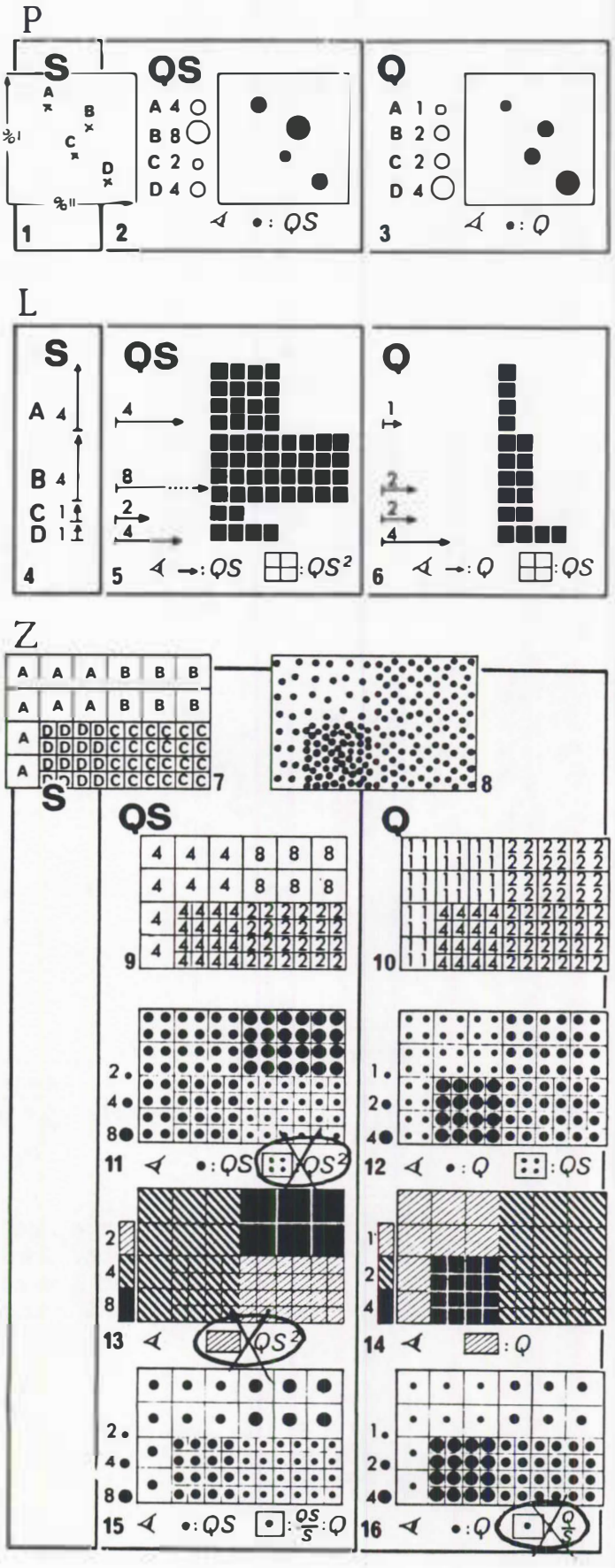
Classes (communes)	A	B	C	D	
Surfaces (S)	4	4	1	1	(dizaines de km²)
Quantité de populat. (QS)	4	8	2	4	(milliers de personnes)
Densité de populat. (Q)	1	2	2	4	(00)

En (1) les communes sont en **implantation ponctuelle.** Ce sont les points d'un diagramme de corrélation (répartition des communes suivant le % de population agricole (I) et industrielle (II).
Chaque point peut recevoir en 3e dimension soit des quantités QS (2), soit des quantités Q (3) que l'œil percevra correctement.

En (4) les communes sont en **implantation linéaire** verticale et proportionnelle à S. Si l'on construit les quantités QS sur l'autre dimension du plan (5), l'œil perçoit horizontalement les QS, mais il voit surtout la surface construite, c'est-à-dire QS², qu'il interprète comme étant la population QS. Surfaces et profil sont erronés. Il faut donc construire horizontalement QS/S, c'est-à-dire Q (6) qui donne, en surface, une image exacte de la quantité QS, et horizontalement une image exacte de la densité Q.

En (7) les communes sont en **implantation zonale** proportionnelle à S. Les QS et les Q se distribuent suivant (9) et (10). La représentation la plus simple (1 point pour 1.000 habitants) fournit l'image (8) qui est incontestable.
On juge aisément des confusions visuelles auxquelles aboutissent les constructions (11) et (13) qui étendent la valeur QS à toute la zone.
L'œil y voit, comme en (5), QS multiplié par la surface, c'est-à-dire QS² (voir aussi p. 76, 5 et 6).
La représentation des QS peut cependant être utile (c'est, par exemple, la mesure de la responsabilité des maires). Dans ce cas, la construction (15), c'est-à-dire un point QS par zone, évite les confusions visuelles précédentes.
Par contre, construire un point Q par zone conduit à une représentation erronée (16).

Il est intéressant de constater que l'erreur de perception fournie par la construction (5) est bien connue des statisticiens et qu'elle est toujours évitée, tandis que les perceptions erronées fournies par les images (11) et (13), et dont l'erreur s'exprime mathématiquement de la même manière (perception de QS²) se rencontrent encore. Le contrôle de la perception en troisième dimension est moins évident que le contrôle de la perception dans le plan. Il n'en est pas moins important puisqu'il intéresse toute la cartographie.



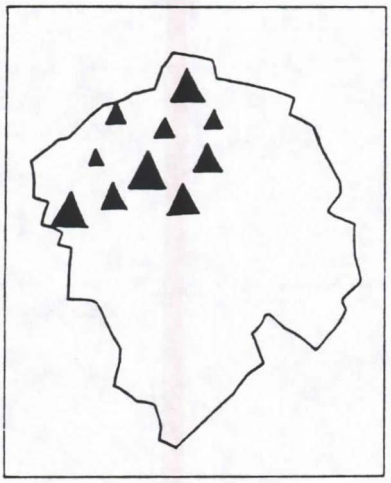
2. Le plan est homogène et continu

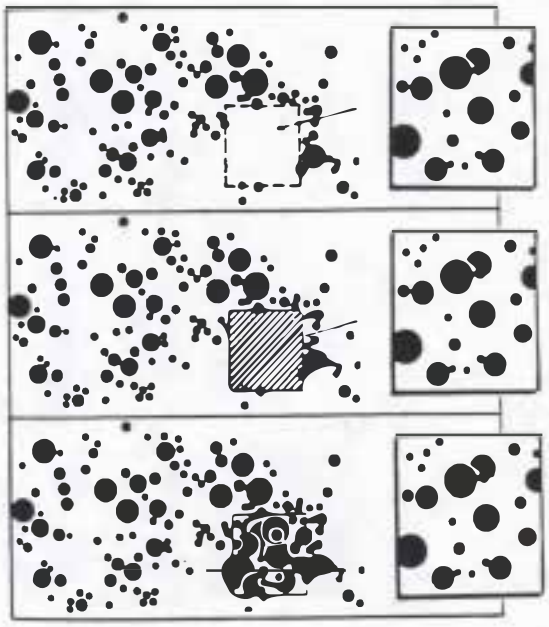
a) Un même signe a une même signification quelle que soit sa position (une convention est invariable).

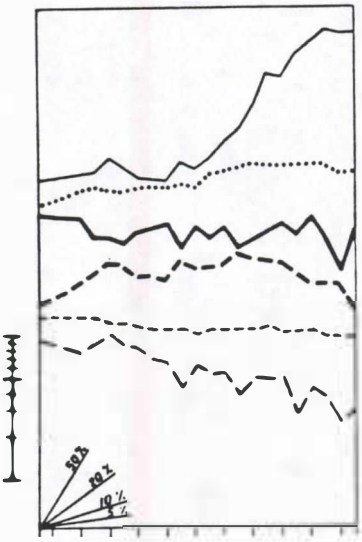
b) L'absence de signe signifie l'absence de phénomène.

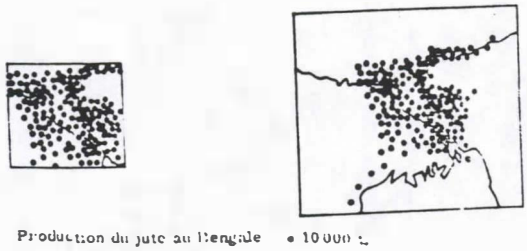
c) Toute variation visuelle apparaît comme signifiante.

d) Le cadre limite le plan significatif.









Production du jute au Bengale • 10 000 t.

.../...

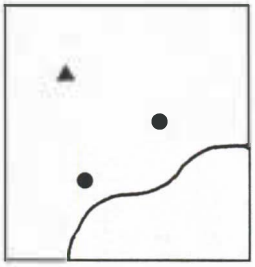
3. Le niveau d'organisation du plan

Le niveau d'organisation fait appel à trois potentialités pour un système. Il est

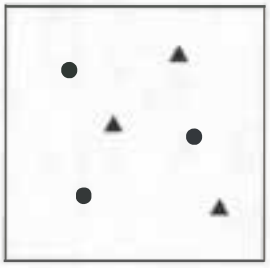
sélectif	associatif	ordonné	et/ou	quantitatif
≠	≡	O		Q

ainsi, dans le plan, les variations de position sont

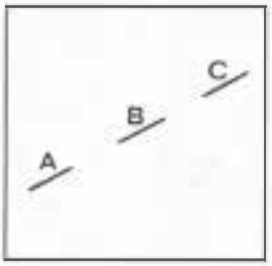
sélectives, puisqu'elles permettent de différencier des objets



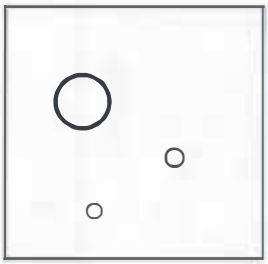
associatives, dans la mesure où deux figures semblables représentent le même objet



ordonnées (les figures A, B et C ne seront jamais lues A C B)



quantitatives, puisque les propriétés métriques du plan permettent d'induire des relations quantifiables.



Il est clair que le niveau d'organisation du plan est associé au niveau d'organisation des variables qui sont elles-mêmes sélectives, associatives, ordonnées et/ou quantitatives comme nous le verrons plus loin.

.../...

B - LES VARIABLES

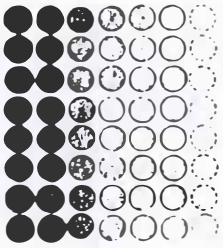
La taille la valeur le grain la couleur l'orientation la forme

1. Les perceptions

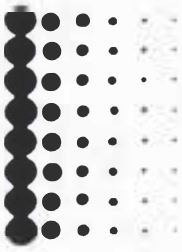
a) Associative ≡

La perception associative est utilisée pour établir des correspondances. Une variable associative ne saurait faire varier la visibilité des signes ; pour cette raison, **la valeur et la taille ne sont pas associatives.**

VALEUR ≠



TAILLE ≠

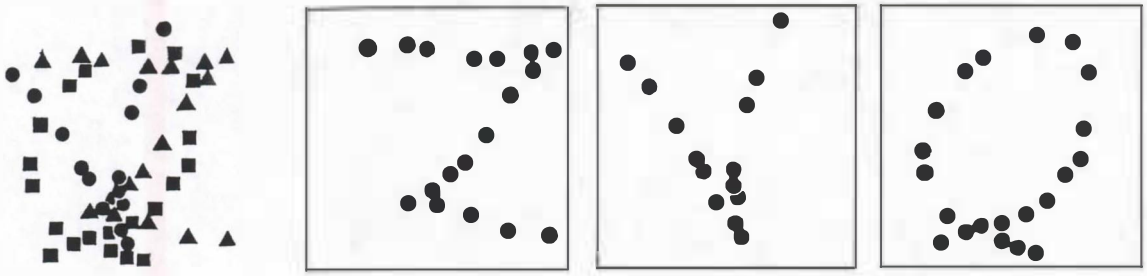


Lorsqu'on s'éloigne, on constate que certains signes disparaissent.

La forme, l'orientation, la couleur et le grain sont associatives.

b) Sélective ≠

On recherche des variables qui permettent d'isoler des objets de même nature et d'en caractériser les structures (on va donc plus loin qu'avec la perception "associative"). La forme n'est pas sélective et l'orientation n'est sélective que pour les implantations points et lignes ; les autres variables sont sélectives.



.../...

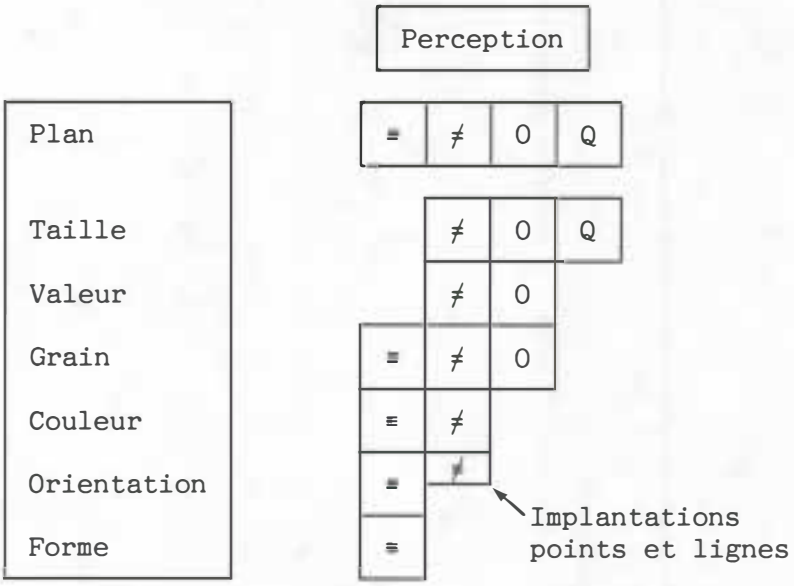
c) **Ordonnée O**

On cherche à comparer plusieurs séries d'objets. Les formes, les orientations, les couleurs ne sont pas (ou peu) ordonnées. Il en va différemment des grains, des valeurs et des tailles.

d) **Quantitative Q**

Seule la taille est ordonnée (la valeur ne l'est pas en raison de l'ambiguïté sur l'évaluation du blanc.

RESUME



.../...

2. Propriétés des variables

a) Taille

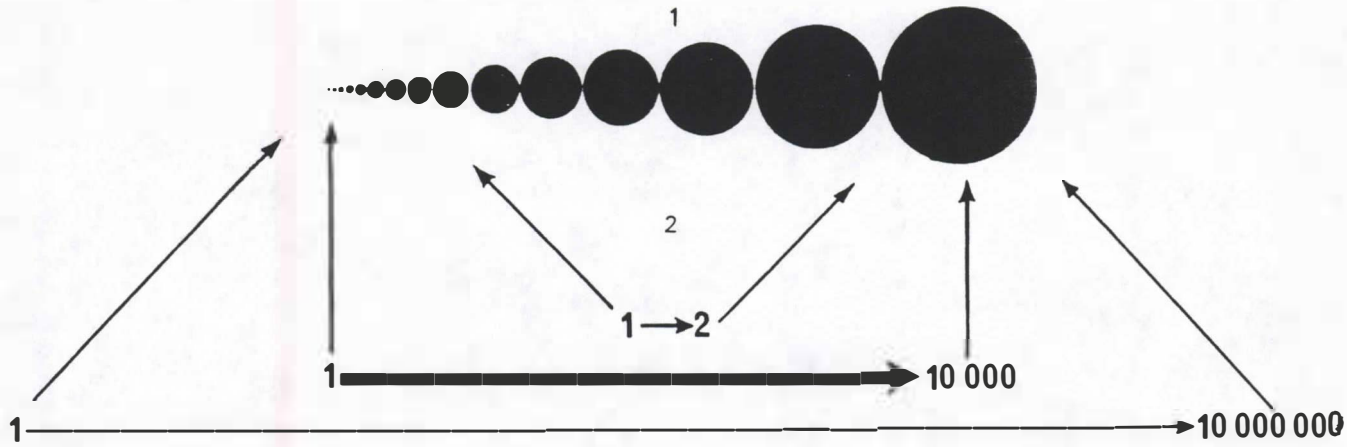
En implantation ponctuelle, on dispose de points dont l'un a une surface 10 000 fois plus grande que l'autre.

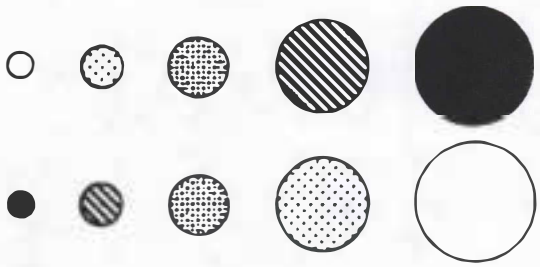
L'oeil différencie en moyenne vingt paliers distribués entre des surfaces allant de 1 à 10.

Cependant, en perception sélective, la variation de taille utilisable est courte (4 à 5 paliers au plus).

Rappel : La variable taille est dissociative ; c'est la seule qui permette de quantifier.

Associée à d'autres variables, elle est généralement dominante mais son interaction avec la variable valeur est assez particulière.





.../...

b) Valeur

Le noir a toujours un impact plus fort que le blanc.

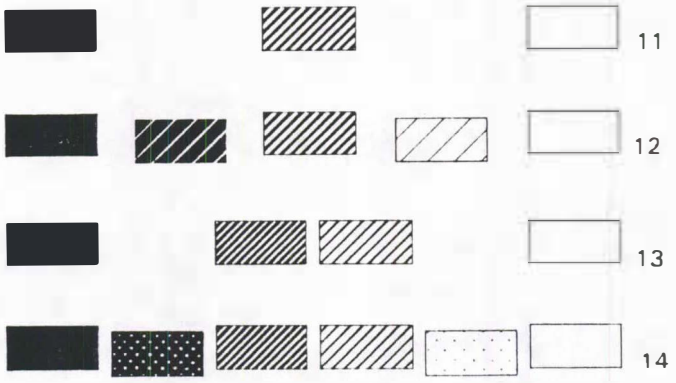
En perception sélective, ne pas utiliser plus de six à sept paliers de gris.

Plus les taches sont petites, plus le nombre de paliers est réduit.

Pour un nombre réduit de paliers, on peut utiliser

. 3 paliers : le noir, le blanc, un gris moyen (plutôt inférieur que supérieur à 50 %)

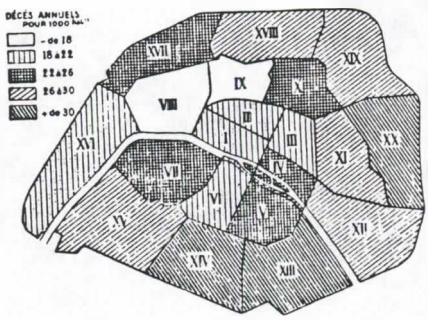
. 5 paliers : les mêmes, plus un noir "rompu", un blanc "rompu".

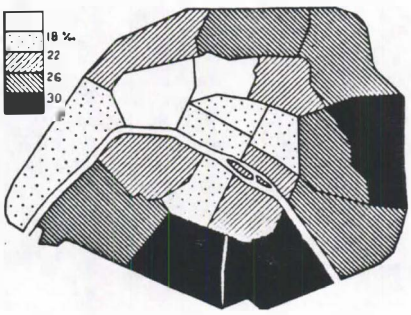


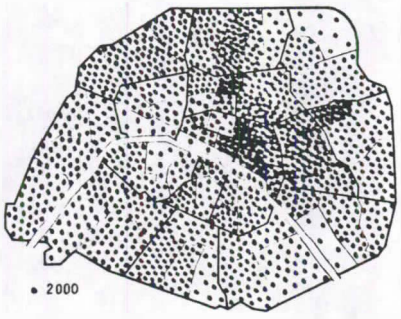
c) Grain

C'est surtout en implantation zonale qu'il peut fournir le maximum de paliers sensibles.

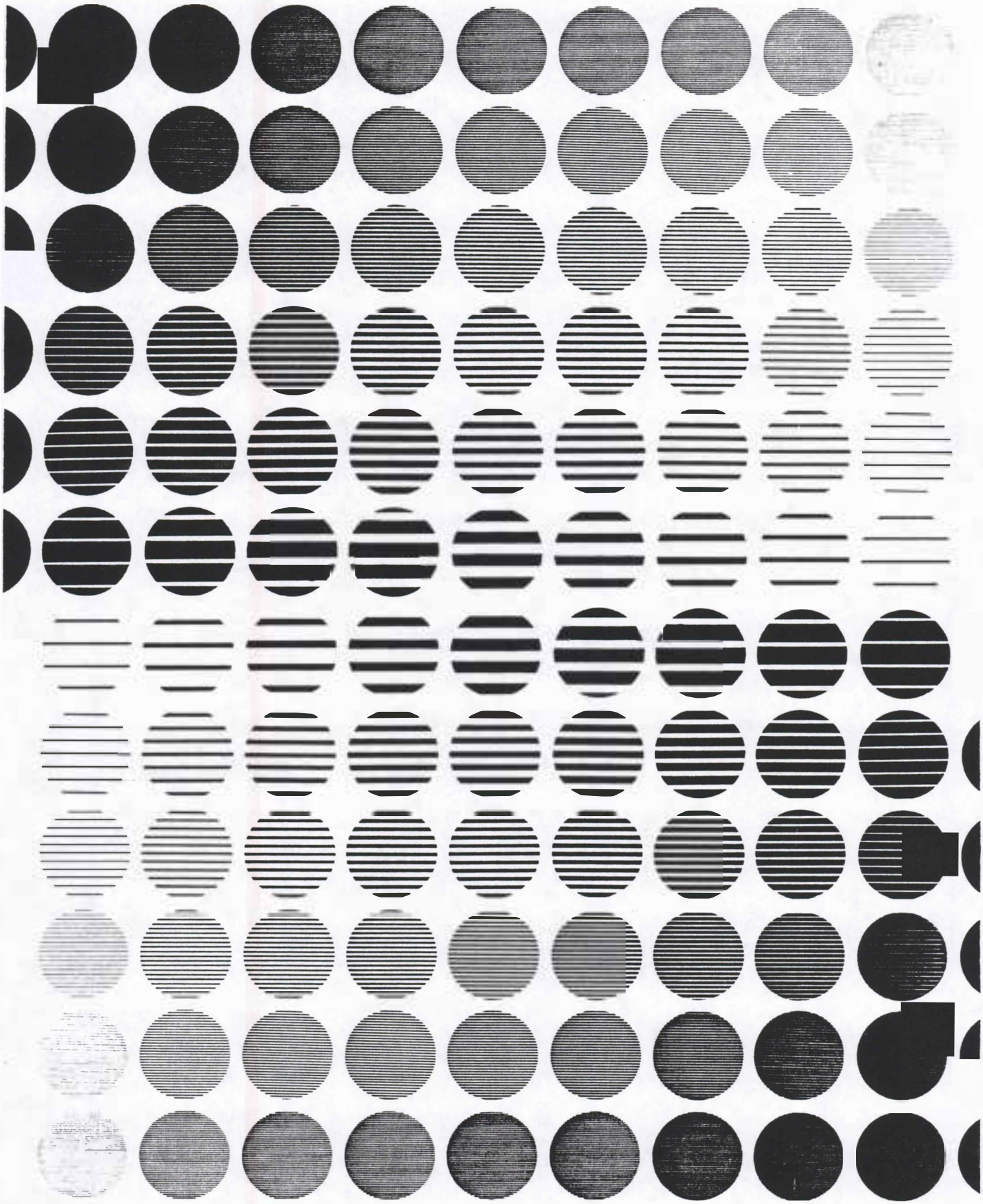
Attention aux effets vibratoires qui provoquent des sensations pénibles et des interprétations ambiguës.







.../...



d) Couleur

Partant d'une couleur, il est possible de la faire évoluer en valeur en ajoutant du noir ou du blanc. Il existe pour chaque couleur une valeur "centrale", dite "saturée" ou "pure" (cette couleur correspond, en fait, à une bande très étroite du spectre coloré).

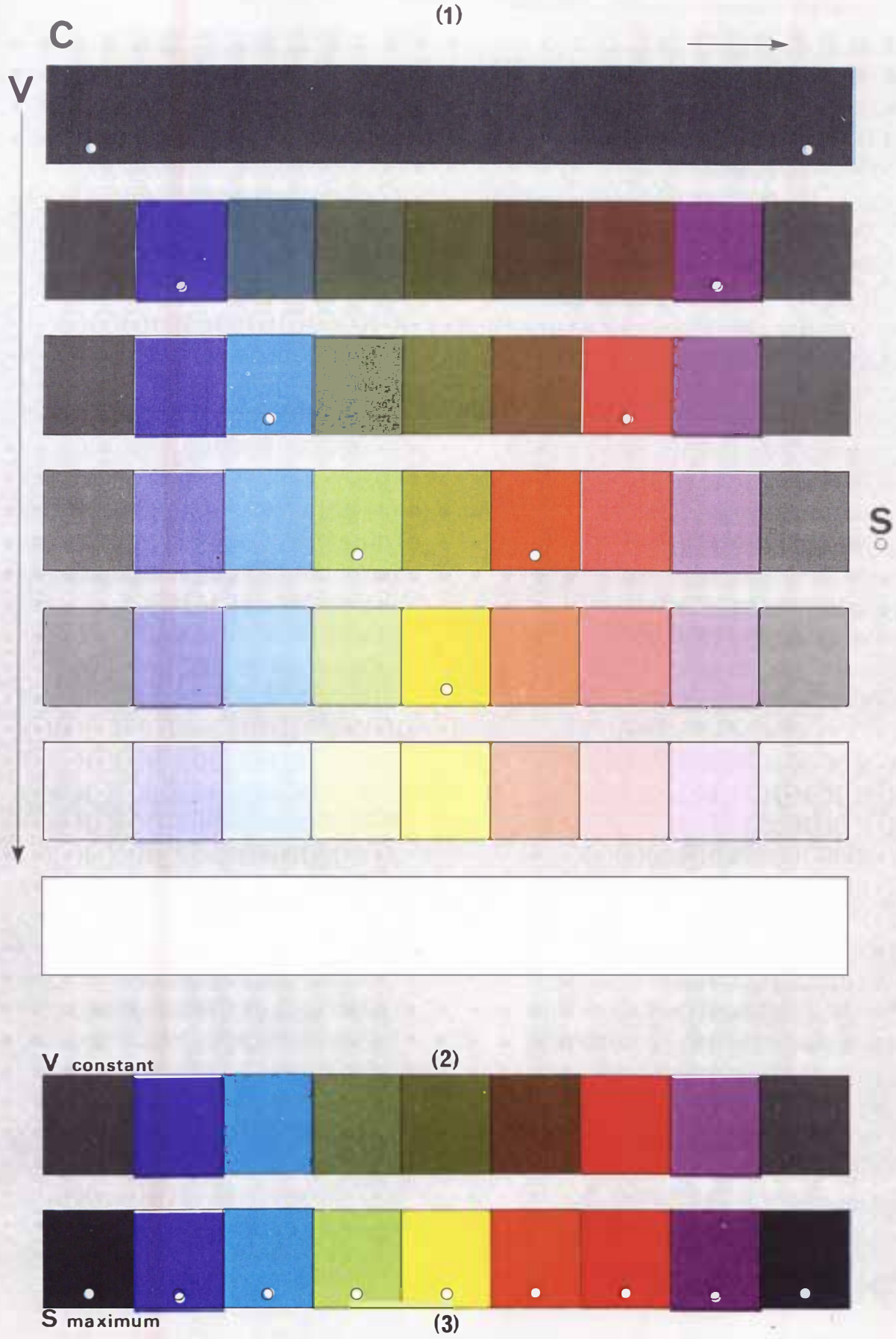
Suivant la couleur, les tout purs varient en valeur, ce qui rend difficile leur emploi. On distingue généralement les couleurs froides et les couleurs chaudes et il est déconseillé de les mélanger.

Remarque : La variation de couleur est inutile à la perception ordonnée.

A valeur égale : La variation de couleur n'est pas ordonnée. Le choix des couleurs sélectives est différent selon la valeur retenue (la sélectivité est maximum près de la couleur saturée).

Avantages	Inconvénients
Perception sélective	Anomalies de la perception
Attraction psychologique	Diffusion coûteuse
Messages pédagogiques	Valeur symbolique difficile à maîtriser

.../...


(1)
C
V
S
V constant
(2)
S maximum
(3)

e) Orientation

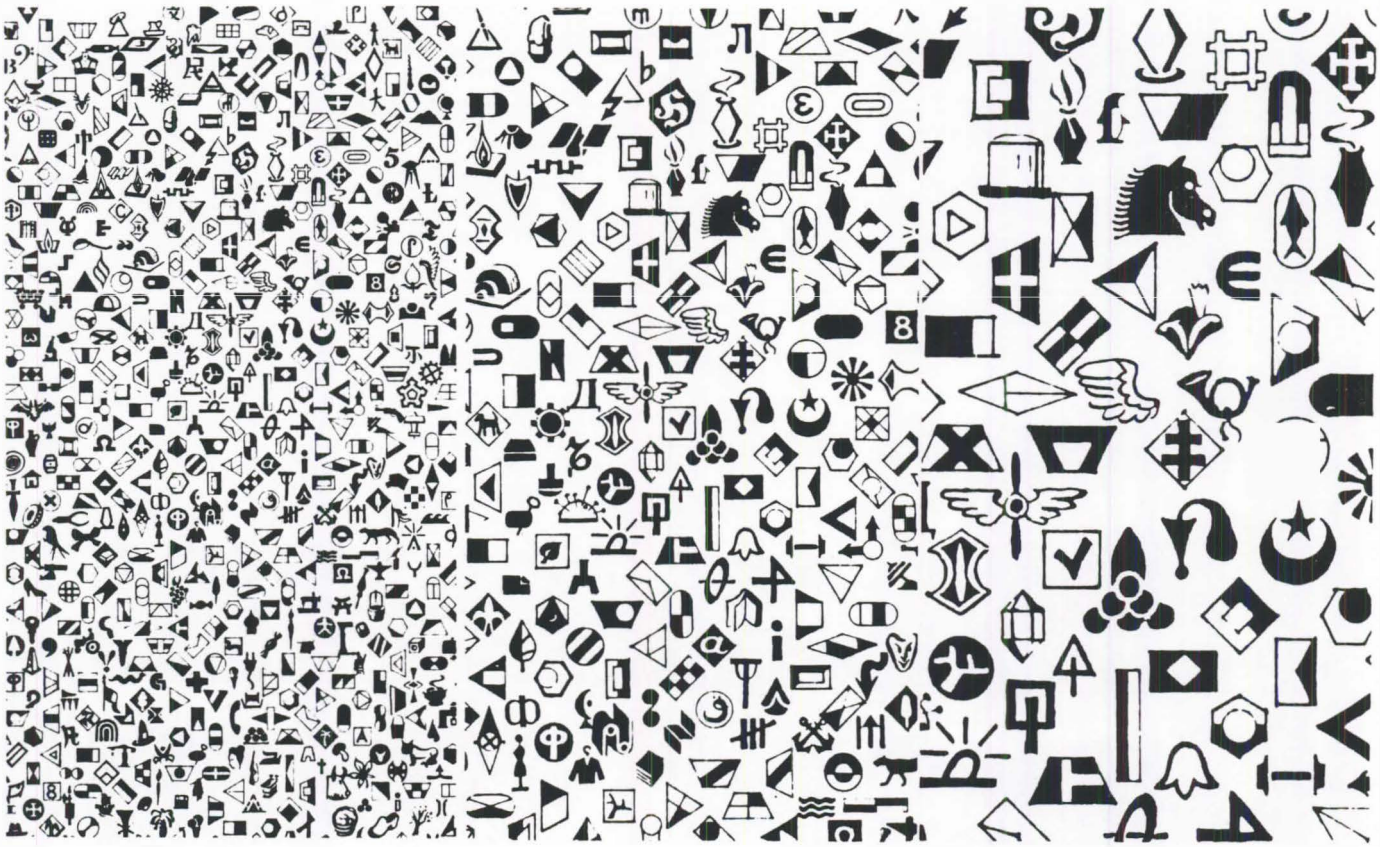
Les signes obliques ont tendance à former une famille particulière par rapport aux horizontales et aux verticales, et ces dernières à "s'opposer".

La perception est sensible à la position de l'observateur.

Ne pas utiliser plus de cinq directions.

f) Forme

Une forme est souvent une représentation signifiante...



REPRESENTATION D'UN CARACTERE QUALITATIF

Septembre 1992

Jean-Claude BERGONZINI

BIOMETRIE
CIRAD - Forêt

BIOM/JCB/AG

REPRESENTATION D'UN CARACTERE QUALITATIF

Fréquemment, on représente les caractères qualitatifs au moyen de camemberts ou secteurs circulaires et de tuyaux d'orgue.

1. Représentation d'un caractère qualitatif

Camemberts : Chaque secteur correspond à une modalité et la fréquence de la modalité ($f_i = \frac{n_i}{n}$) est proportionnelle à l'angle au centre θ_i ainsi qu'à la surface δ_i du secteur

$$\theta_i = 360^\circ \frac{n_i}{n} \qquad S_i = 180^\circ \frac{n_i}{n} R^2$$

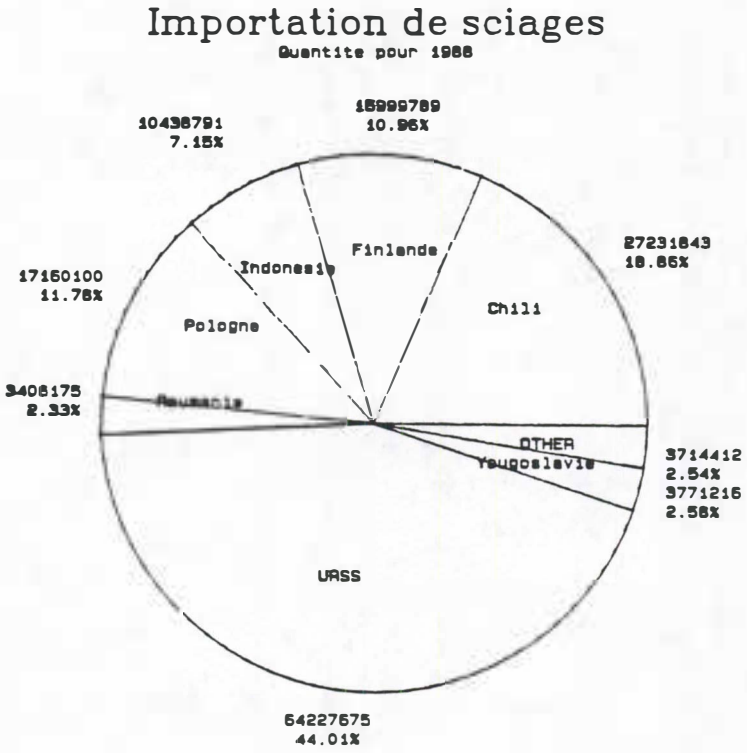
Tuyaux d'orgue : La hauteur est proportionnelle à la fréquence. Tous les tuyaux ont même longueur. La surface du tuyau est proportionnelle à f_i .

Voir illustrations page suivante.

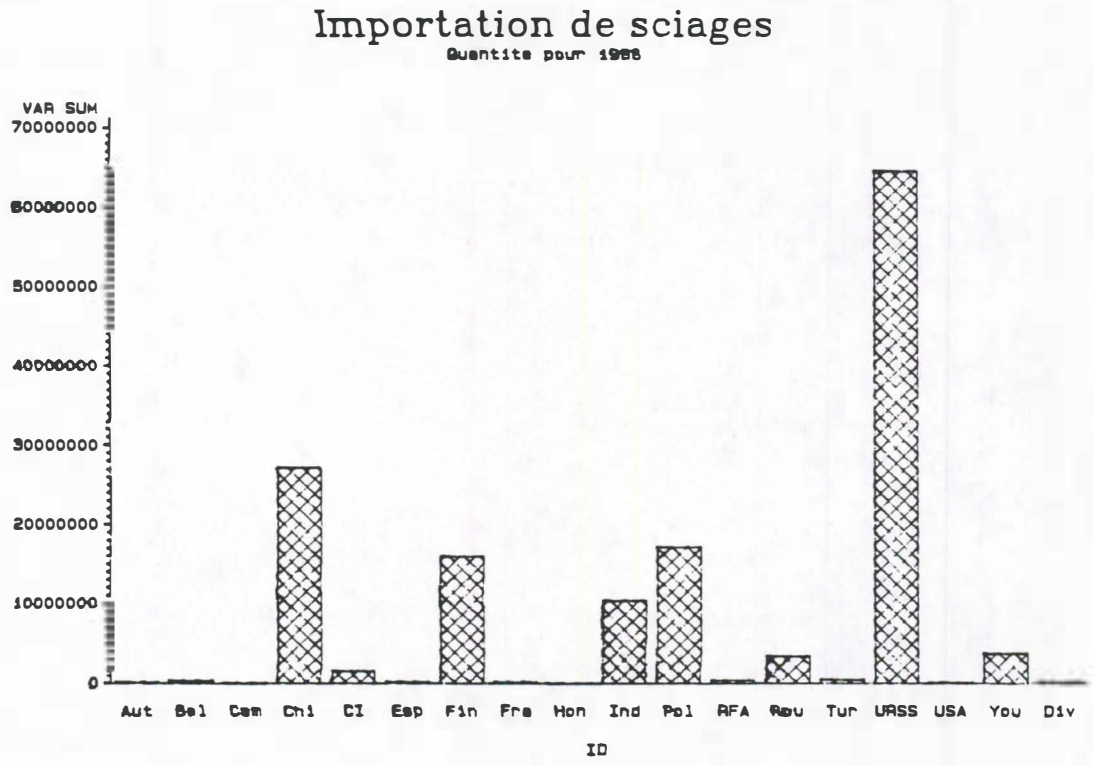
Remarque : Les camemberts offrent une perception directe des fréquences.

.../...

Exemple de camembert



Exemple de tuyaux d'orgue



.../...

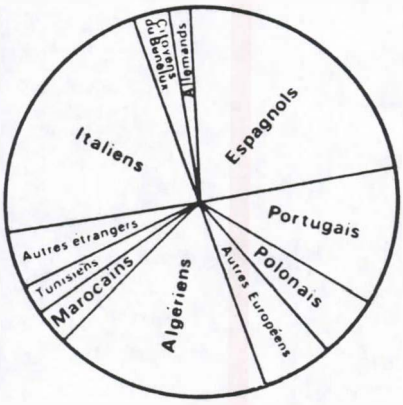
2. Limites et amélioration

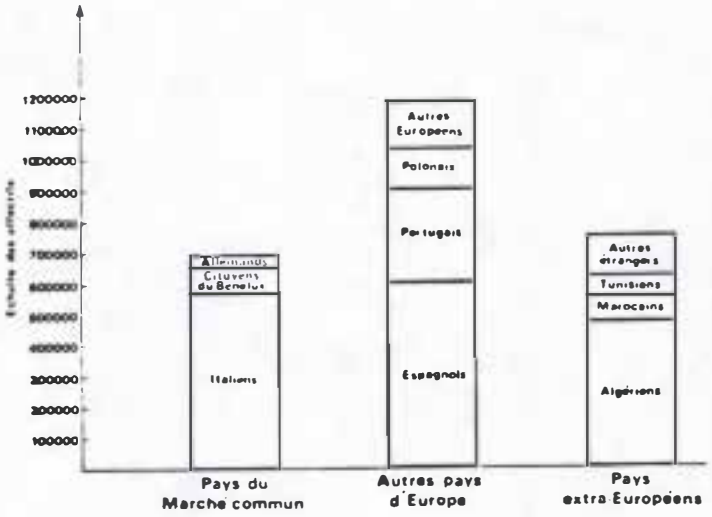
Le facteur limitant (principalement pour les camemberts) est l'accroissement du nombre de modalités ou le déséquilibre entre leurs fréquences. Les solutions les plus immédiates sont :

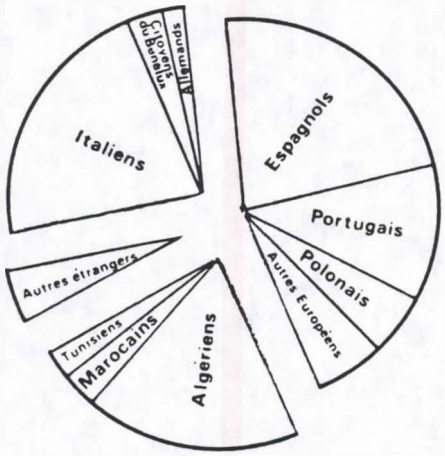
- l'augmentation de R
- le regroupement de certaines modalités et l'excentrage des secteurs
- l'utilisation de la couleur ou de trames.

Attention toutefois :

- R ne peut pas être pris trop grand
- les regroupements doivent correspondre à une véritable logique ; Ils peuvent disperser l'attention et rendre la lecture difficile ;
- l'utilisation de couleurs ou de trames n'est pas toujours une amélioration.







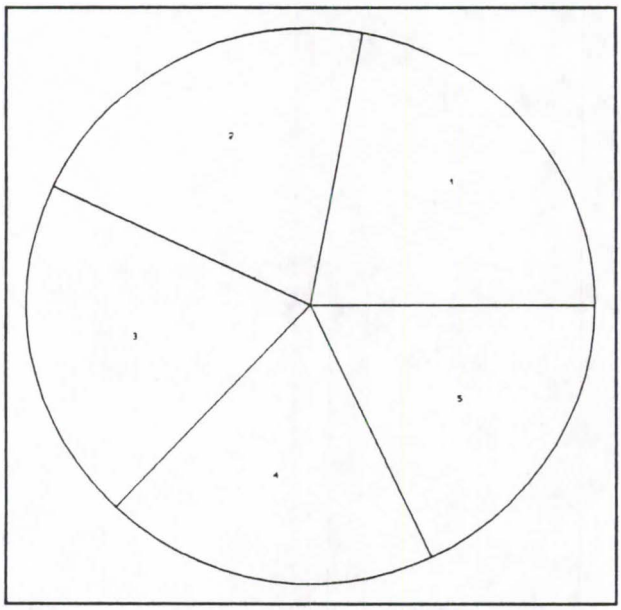
Etrangers résidant en France selon la nationalité (1968)

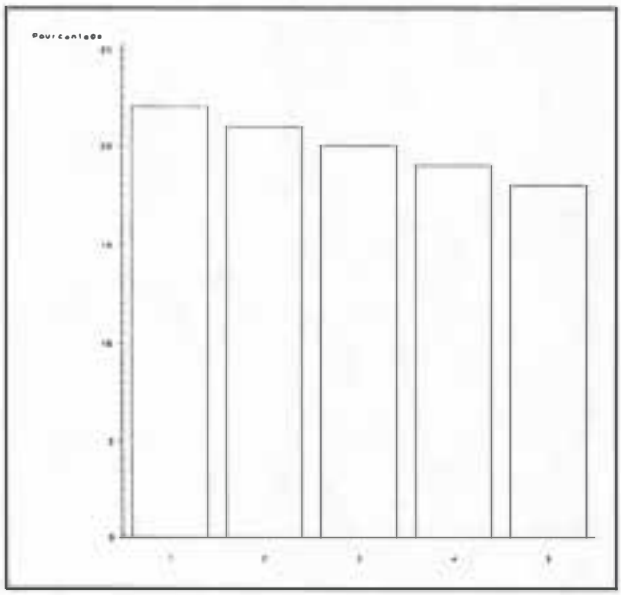
.../...

Jugement de la valeur des modalités :

Lors de la visualisation d'un camembert, il est parfois difficile d'ordonner les valeurs des modalités par ordre croissant, en particulier lorsque ces valeurs ne sont pas très différentes.

Un exemple est présenté, à l'aide d'un camembert et d'un tuyau d'orgue. Les graphiques parlent d'eux-mêmes.





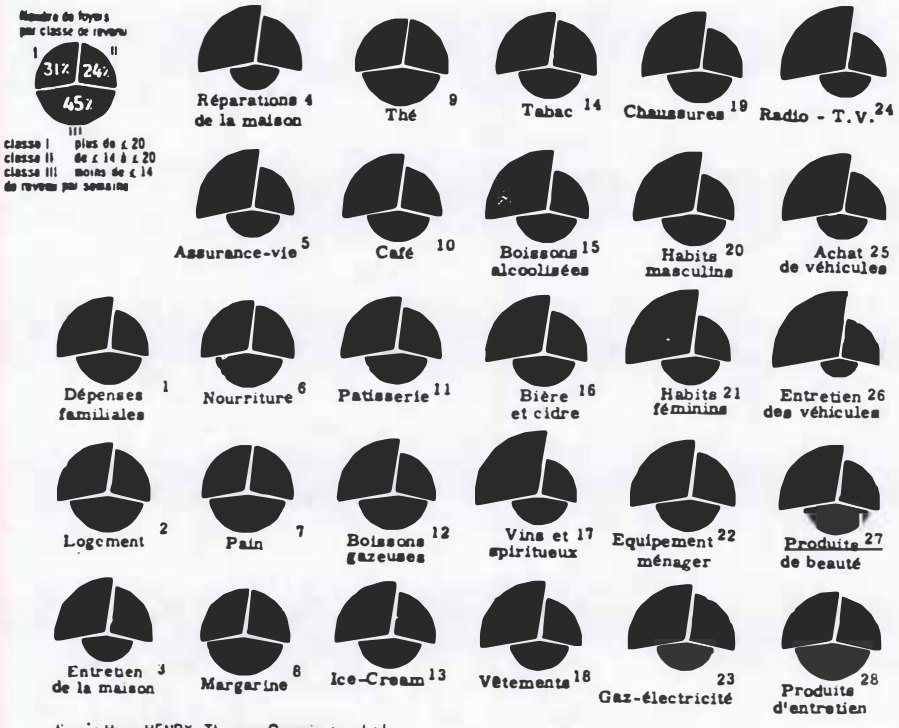
.../...

3. Comparaison de populations

L'utilisation de tuyaux d'orgue est généralement conseillée. Toutefois, dans le cas où les populations à comparer sont associées à une répartition spatiale, il est fréquent d'utiliser des camemberts.

Fig. 1 : Plusieurs images à mémoriser - sentiment d'uniformité.

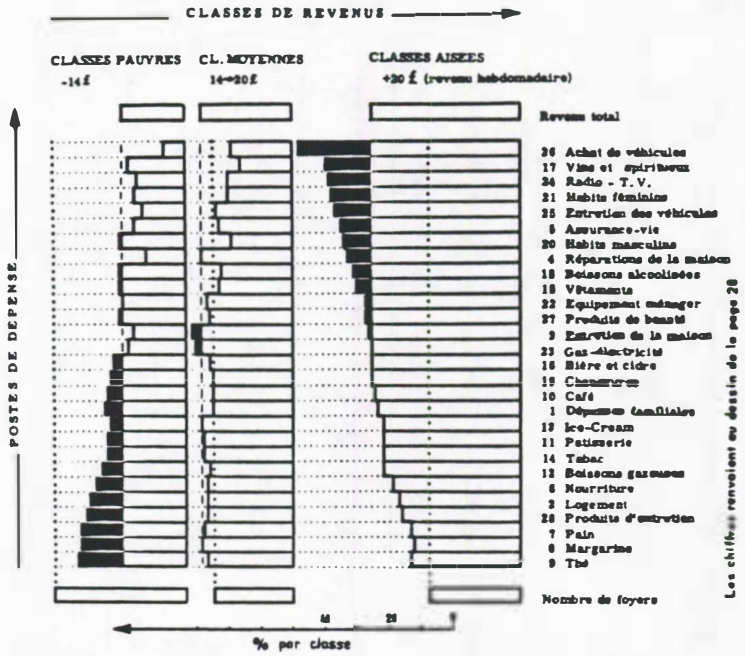
COMPARAISON DES POSTES DE DEPENSE
SUIVANT LES CLASSES DE POPULATION
DANS LE ROYAUME-UNI . EN 1960.



d'après Harry HENRY. Thomson Organisation Ltd.
Sources Central Statistical Office. London 1961

.../...

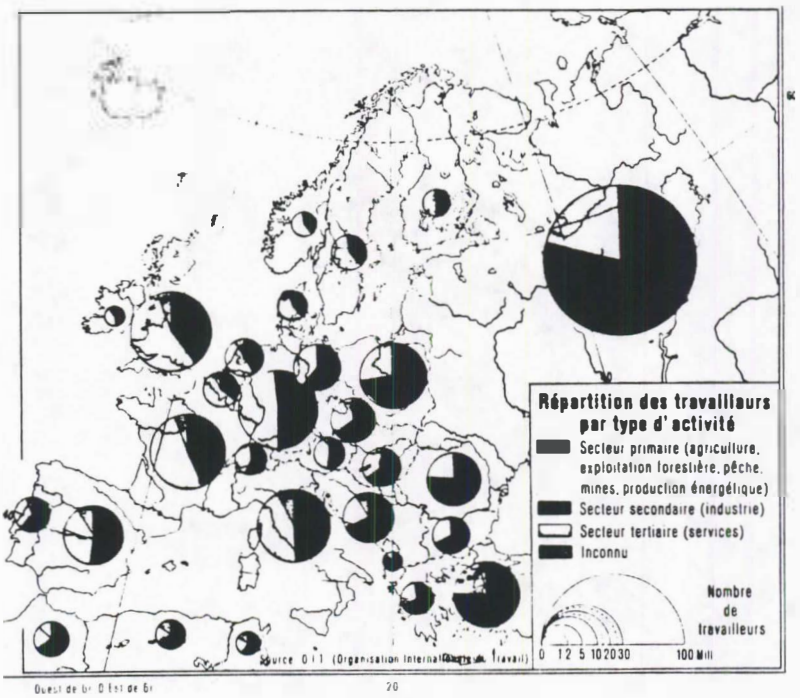
Fig. 2 : Une image globale mémorisable - différences visibles et mieux mises en évidence.

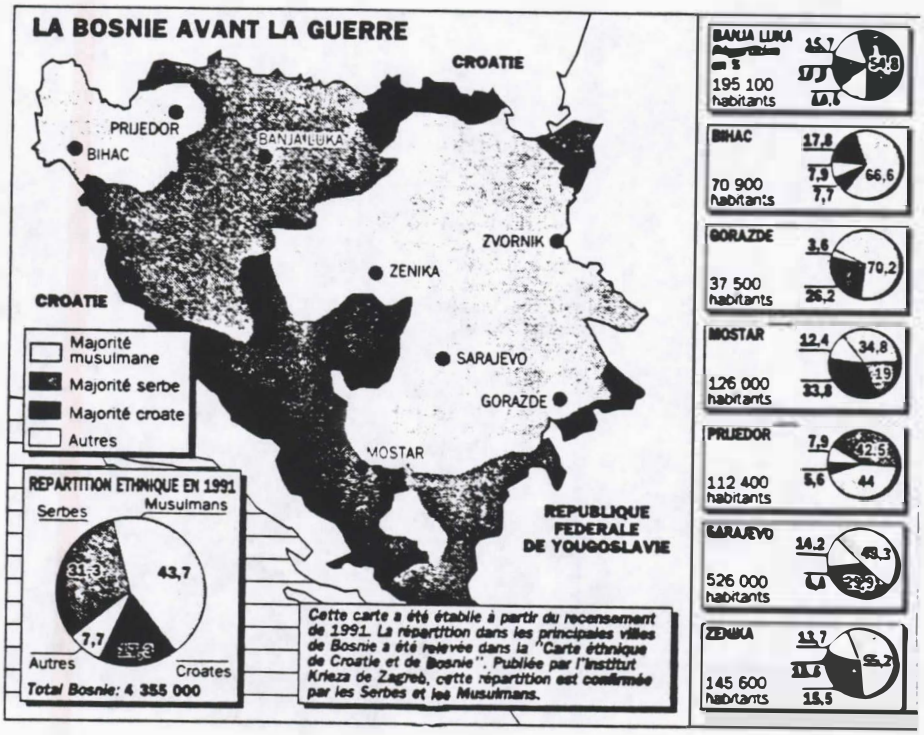


4. Introduction d'une variable supplémentaire

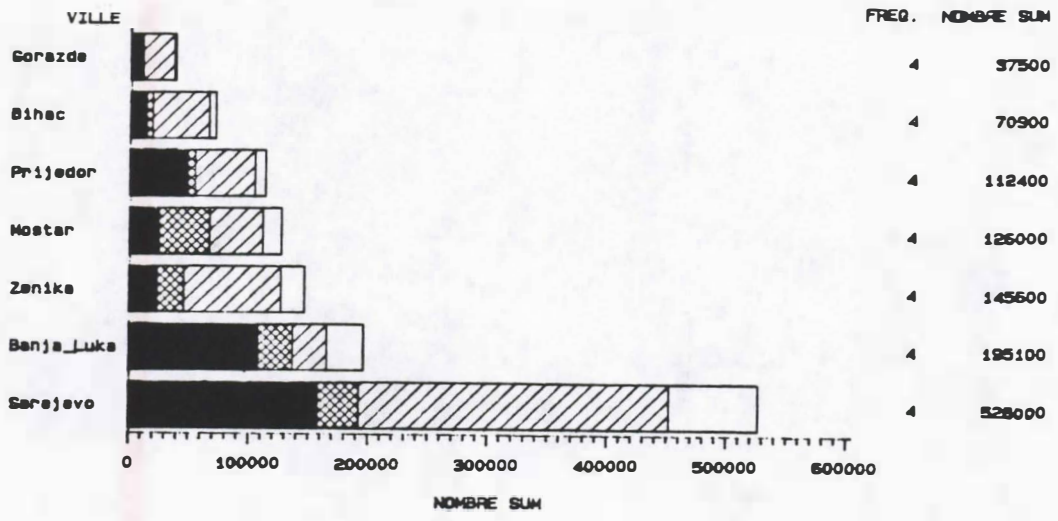
Dans l'utilisation des camemberts, on peut faire varier le rayon proportionnellement à une variable quantitative.

Les tuyaux d'orgue semblent mieux adaptés à l'insertion de variables qualitatives.





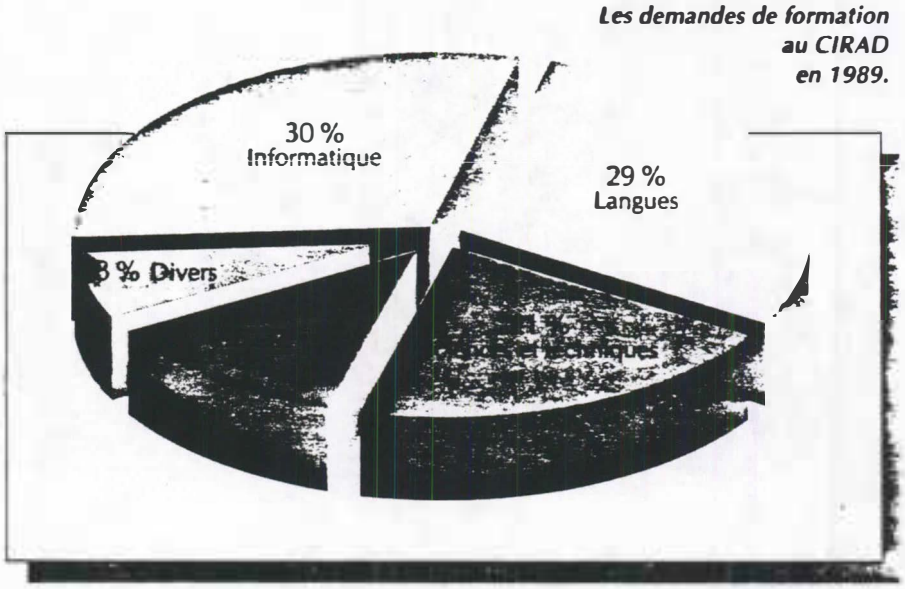
Repartition des communautes dans 7 villes de YOUGOSLAVIE



.../...

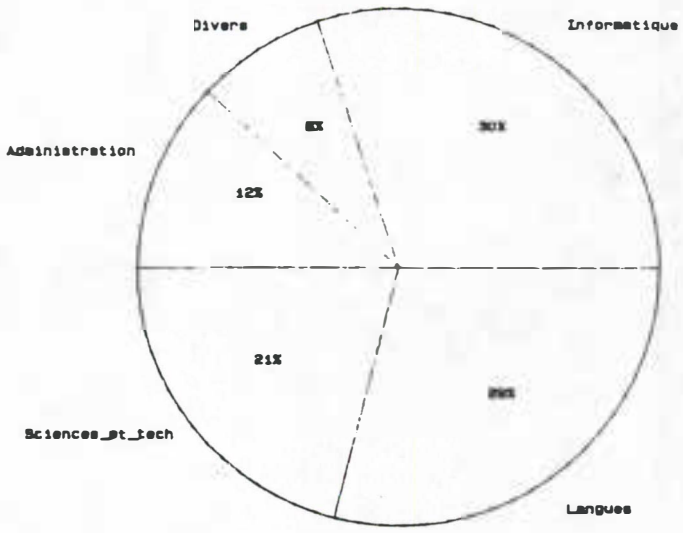
5. Esthétisme et objectivité

L'esthétisme, même s'il ne déforme pas de façon sensible l'information, ne respecte pas toujours les quantifications d'origine.



Les demandes de formation au CIRAD en 1989.

Les demandes de formation au CIRAD en 1989



LES HISTOGRAMMES

Septembre 1992

Hervé LEDOUX

BIOMETRIE
CIRAD - Forêt

Les Histogrammes

Les Histogrammes

1. Introduction :

L'histogramme est un graphique résumant la distribution d'une variable continue. Il est l'analogue des diagrammes en bâtons pour les variables discrètes. Cette représentation permet de visualiser la symétrie de la distribution, son étendue, et son allure générale.

Nous présenterons quelques graphes pouvant être considérés comme les ancêtres des histogrammes. Nous détaillerons la construction d'un histogramme, la détermination du nombre de classes et la largeur de celles-ci. Nous terminerons par la présentation d'un graphique similaire : les polygones de fréquence.

2. Un peu d'histoire :

Le premier diagramme en bâtons connu fût introduit par William PLAYFAIR (1759-1823) dans son ouvrage intitulé "*The commercial and political atlas*" publié à Londres en 1786. Ce graphe représente les exportations et les importations d'Ecosse vers ou en provenance de différents pays entre Noël 1780 et Noël 1781 (Figure 1).

En 1801, William PLAYFAIR récidive avec le premier graphe représentant les valeurs d'une variable par des surfaces. Ce graphe représente le nombre d'habitants de quelques villes d'Europe (Figure 2).

Figure 1 :

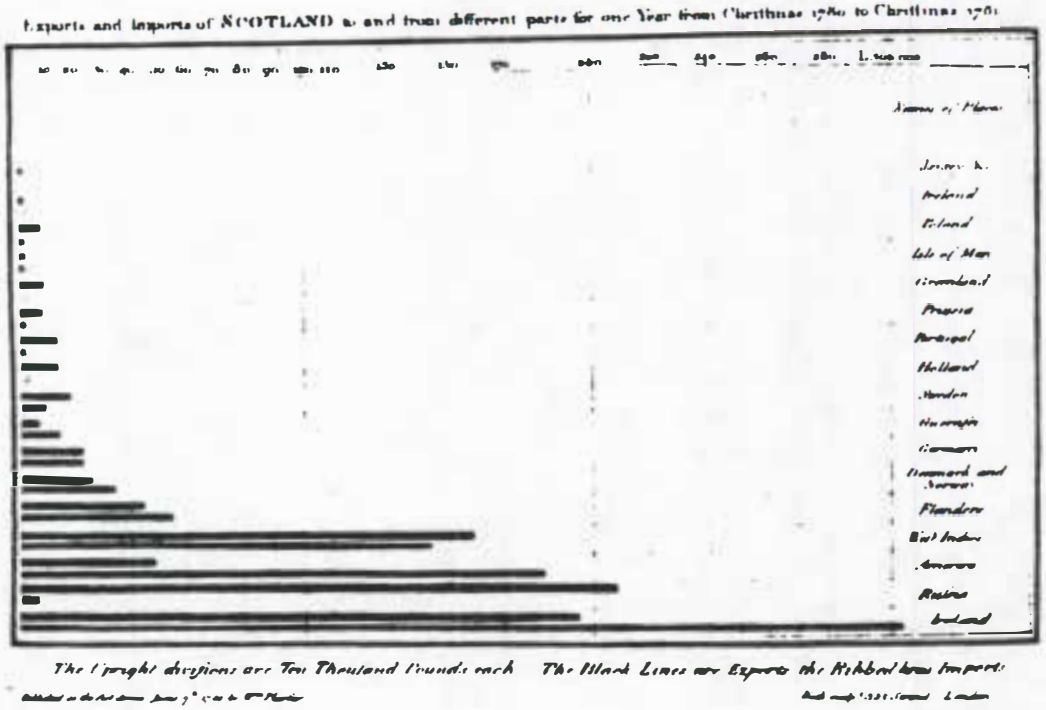
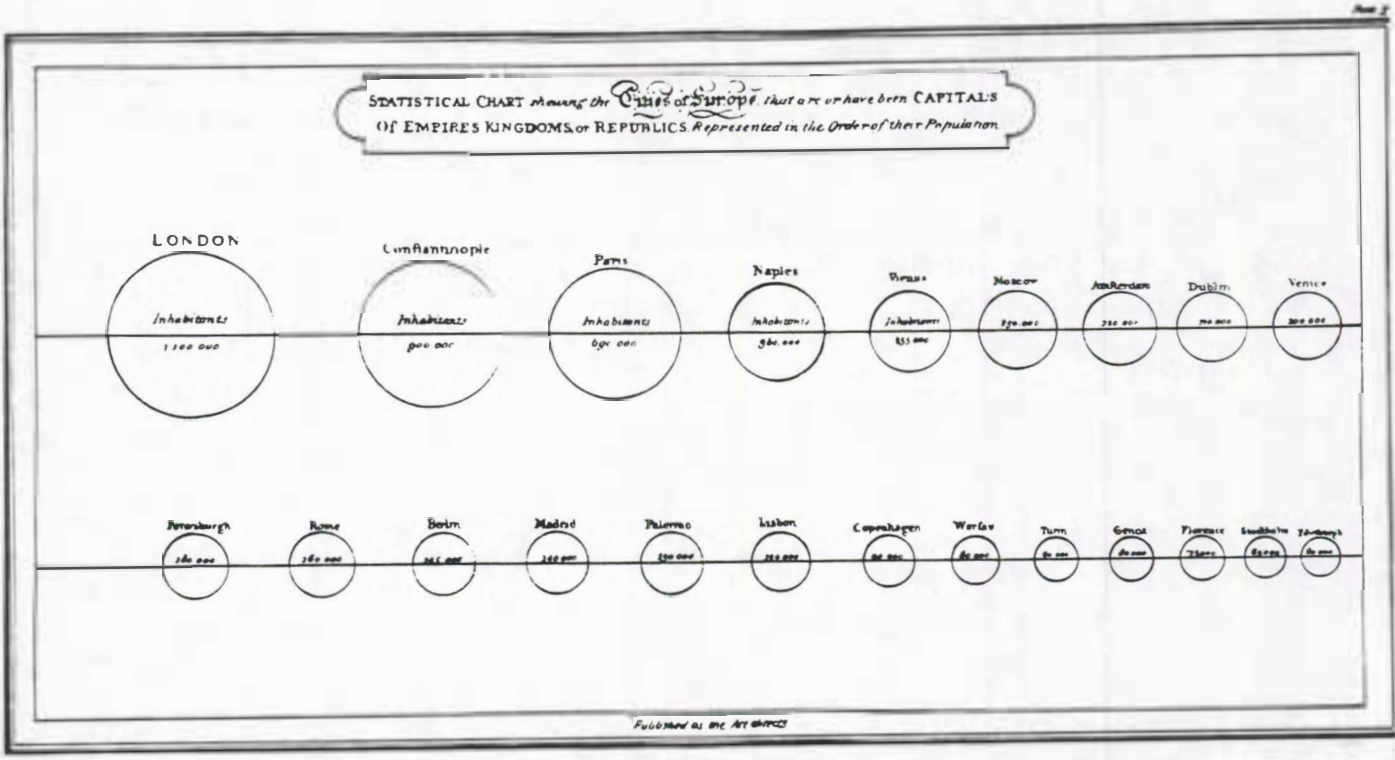


Figure 2 :



3. Construction d'un histogramme :

Pour expliquer la construction d'un histogramme, nous utiliserons un exemple représentant le nombre d'ouvriers par classes de salaire annuel (en milliers d'anciens francs, durant l'année 1952).

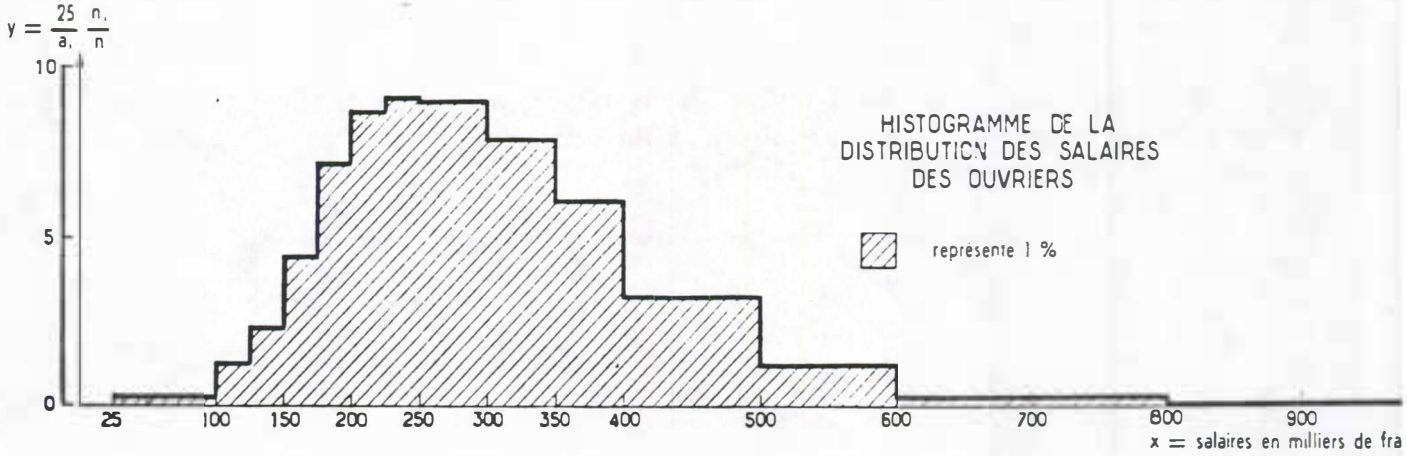
Salaire annuel	Nombre d'ouvriers
Moins de 100	1 721
de 100 à moins de 125	2 413
125 150	4 342
150 175	8 264
175 200	13 300
200 225	16 053
225 250	16 774
250 300	33 251
300 350	29 211
350 400	22 453
400 500	24 005
500 600	9 477
600 800	4 093
800 1 000	443
1 000 1 500	125
1 500 2 000	12
2 000 5 000	14
Total	185 951

Pour chacune des classes, nous calculons la fréquence f_i qui est égale au rapport en pourcentage de l'effectif de la classe sur l'effectif total ; ainsi que l'amplitude de la classe a_i (différence des extrémités de la classe $a_i = e_{i+1} - e_i$).

Extrémités de classe	Effectifs	Fréquences	Amplitudes
125 - 150	4 342	2,34 %	25
150 - 175	8 264	4,44 %	25
...			
350 - 400	22 453	12,07 %	50
400 - 500	24 005	12,91 %	100

L'histogramme est la courbe obtenue en juxtaposant des rectangles dont la base est l'intervalle (e_i , e_{i+1}) et la hauteur f_i/a_i (Figure 3). Ainsi la surface de ce rectangle est-elle proportionnelle à la fréquence f_i.

Figure 3 :



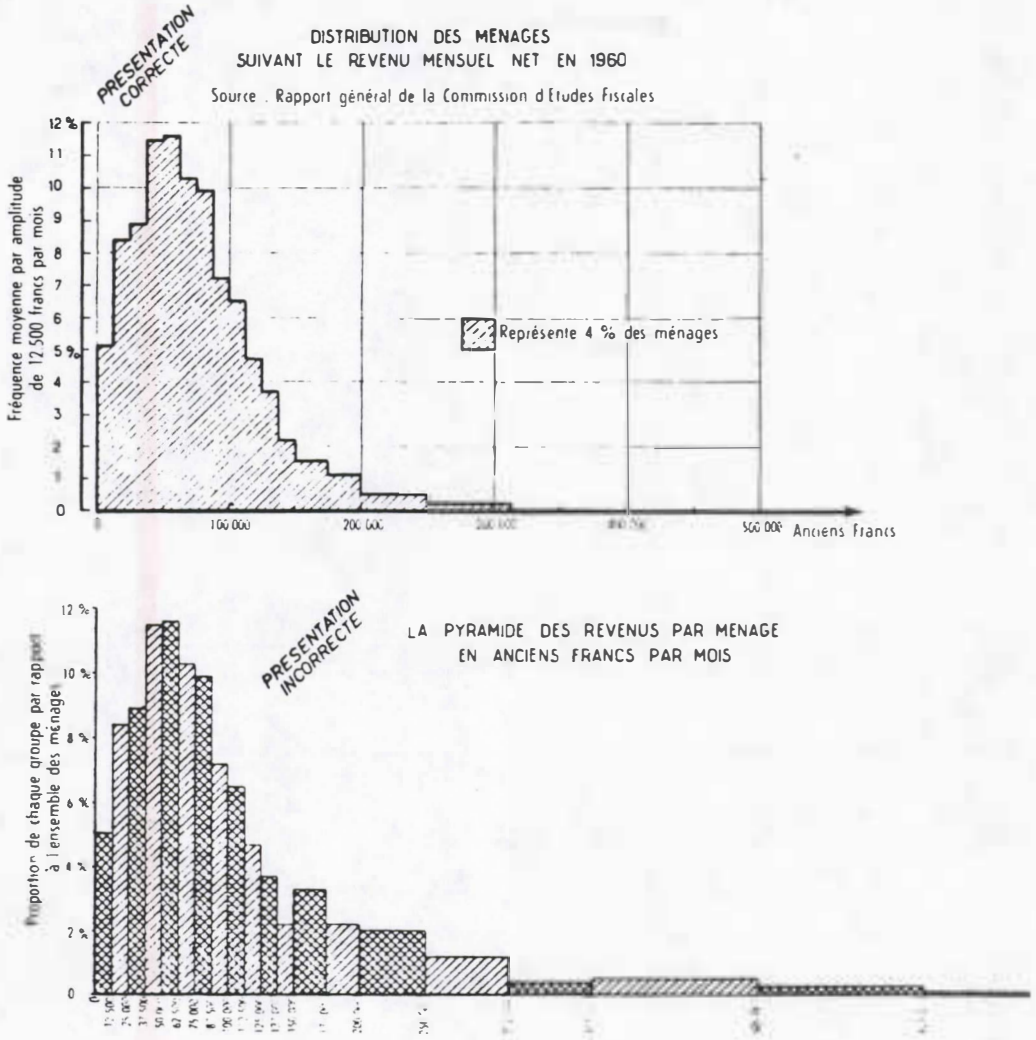
Remarques :

Dans cet exemple, la classe *Moins de 100* a une amplitude indéterminée (au plus égale à 100, puisqu'un salaire est nécessairement positif). Le choix de la valeur de l'extrémité inférieure s'est porté sur 25, puisque la moyenne de cette classe est de 66,2, et que la moyenne la plus proche, de façon à avoir une amplitude multiple de 12,5, est de $62,5 = (100 + 25)/2$.

Une possibilité d'erreur, lors de la construction d'un histogramme, est de dessiner des rectangles dont la hauteur est proportionnelle à f_i et non à f_i/a_i. Dans ce cas la proportionnalité des aires aux effectifs n'est plus respectée. Ce problème ne se pose pas lorsque toutes les classes ont même amplitude.

Un exemple de présentation incorrecte est exposé à la figure 4. Nous apercevons que la classe de revenus 150.000 à 175.000 francs apparaît comme exceptionnelle sur le graphique du bas, alors qu'elle ne l'est en réalité aucunement.

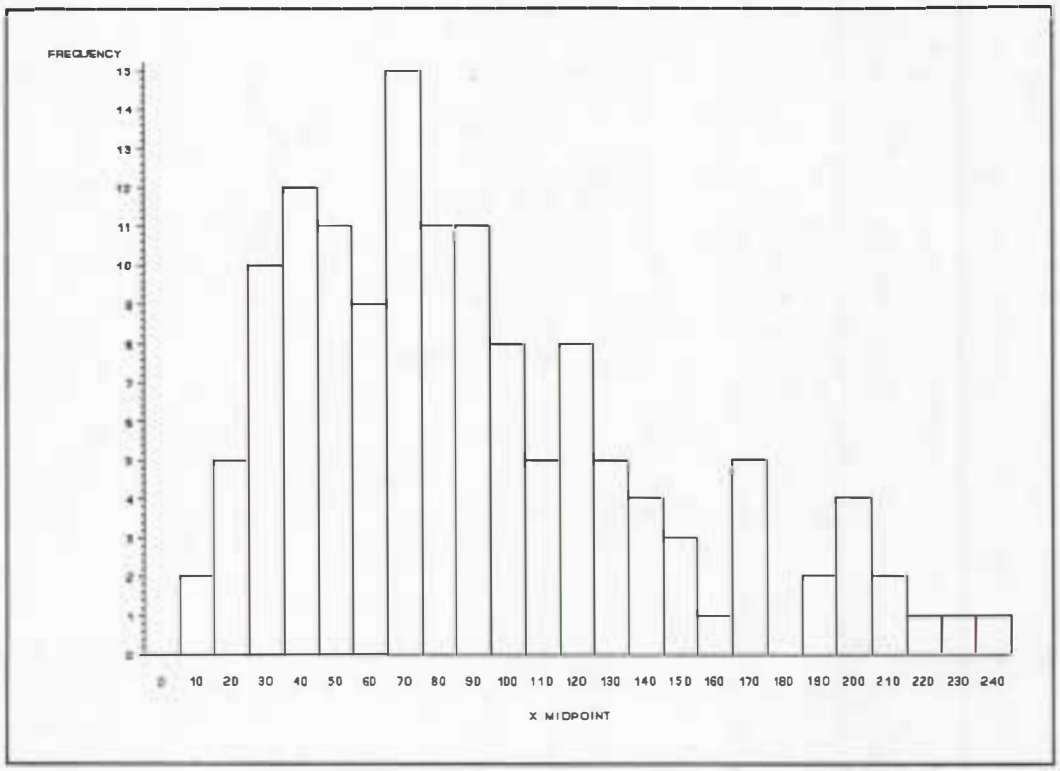
Figure 4 :

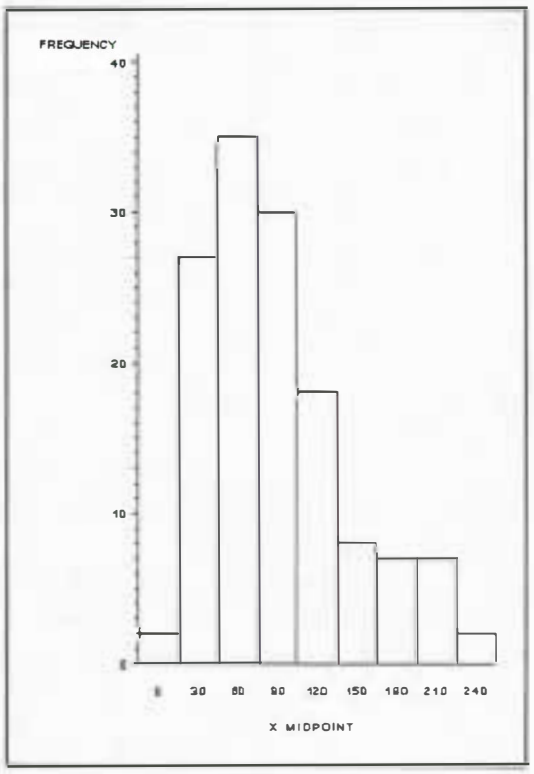


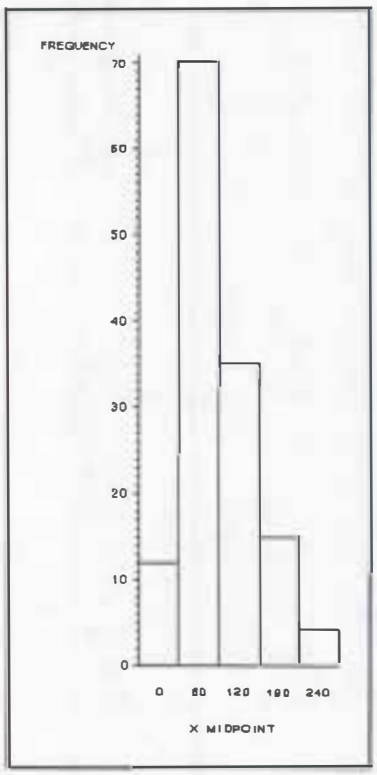
4. Détermination du nombre de classes :

Le nombre de classes est déterminé de façon que la distribution ait une allure régulière sans en dissimuler les caractéristiques essentielles. La figure 5 présente trois histogrammes d'une même distribution, avec trois largeurs de classes différentes égales à 10, 30, et 60. Le premier histogramme est trop irrégulier, nous ne pouvons deviner la distribution sur le troisième, et le deuxième semble être le meilleur compromis entre les deux.

Figure 5 :


FREQUENCY
15
14
13
12
11
10
0
10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240
X MIDPOINT


FREQUENCY
40
30
20
10
0
30 60 90 120 150 180 210 240
X MIDPOINT


FREQUENCY
70
60
50
40
30
20
10
0
0 60 120 180 240
X MIDPOINT

De nombreuses règles de détermination du nombre de classes ont été introduites. DIXON et KRONMAL (1965) propose d'utiliser $L = Int[10 * \log_{10}(n)]$ comme valeur supérieure du nombre de classes. L'expérience a prouvé que cette règle semble donner de bons résultats pour un nombre n d'individus compris entre 50 et 300.

Si n est inférieur à 50, VELLEMAN (1976) propose d'utiliser $L = Int[2n^{1/2}]$.

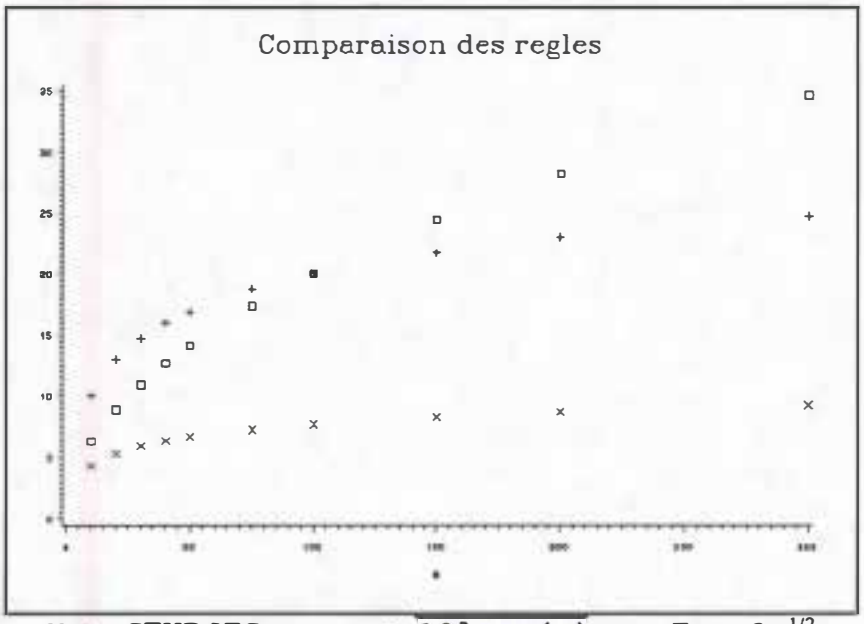
STURGES (1926) a proposé une autre méthode : lorsque n est une puissance de 2, la distribution des fréquences devrait suivre la séquence des coefficients du binôme $C_n^p = \frac{n!}{p!\,(n-p)!}$ (la détermination de ces coefficients est aussi connue sous le nom de triangle de Pascal). Par exemple, pour $n = 16$ individus, la distribution devrait être divisée en 5 classes avec pour fréquence 1, 4, 6, 4, 1. Ce qui est équivalent à prendre $2^{L-1} = n$ ou $L = 1 + \log_2(n)$.

Triangle de Pascal :

1	1					
1	2	1				
1	3	3	1			
1	4	6	4	1		
1	5	10	10	5	1	
1	6	15	20	15	6	1

La figure 6 représente le nombre de classes L calculé suivant les trois règles définies précédemment et fonction du nombre n d'individus.

Figure 6 :



X : STURGES + : $10\log_{10}(n)$ □ : $2n^{1/2}$

La règle de STURGES ne semble appropriée que pour des valeurs de n comprises entre 20 et 40. Les règles $10\log_{10}(n)$ et $2n^{1/2}$ se croisent pour $n = 100$, et il semble qu'il est préférable d'utiliser $2n^{1/2}$ pour n inférieur à 100 et $10\log_{10}(n)$ pour n supérieur à 100.

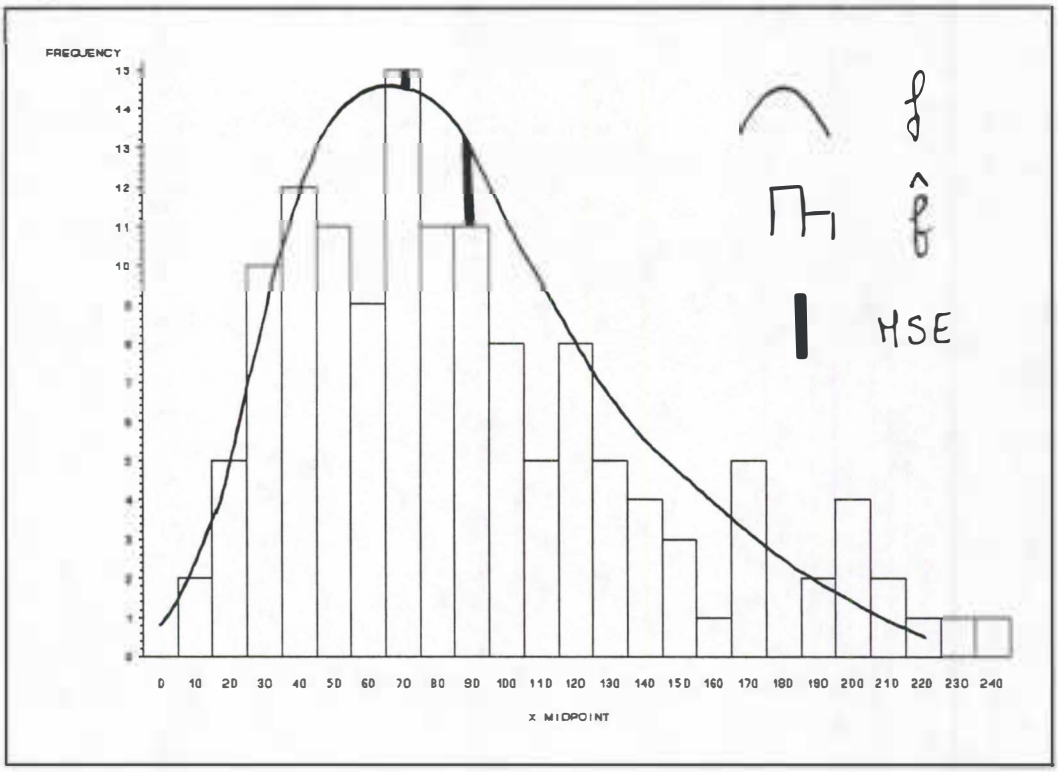
5. Détermination de la largeur des classes :

Un autre moyen pour déterminer le nombre de classes, est de définir la largeur des classes, l'étendue des données étant connue. Cette méthode est plus théorique, et considère l'histogramme comme un estimateur non paramétrique de la densité.

A un point x, l'erreur quadratique moyenne (MSE en anglais) d'un estimateur histogramme $\hat{f}(x)$ d'une densité f(x) est définie par :

$$MSE(x) = E\{\hat{f}(x) - f(x)\}^2$$

Figure 7 :



L'erreur quadratique moyenne intégrée (IMSE) représente l'erreur de mesure globale d'un histogramme et est définie par :

$$IMSE = \int MSE(x)\,dx = \int E\{\hat{f}(x) - f(x)\}^2 dx$$

La largeur des classes h est définie comme minimisant l'IMSE. Si la densité possède deux bornes et une dérivée continue, cette largeur est égale à :

$$h = \left\{ \frac{6}{\int_{-\infty}^{\infty} [f'(x)]^2 dx} \right\}^{1/3} n^{-1/3}$$

Mais la connaissance de la densité f est rare. Par contre si l'on suppose que f représente la densité d'une loi normale de variance σ^2 :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$$

Alors la largeur h s'écrit :

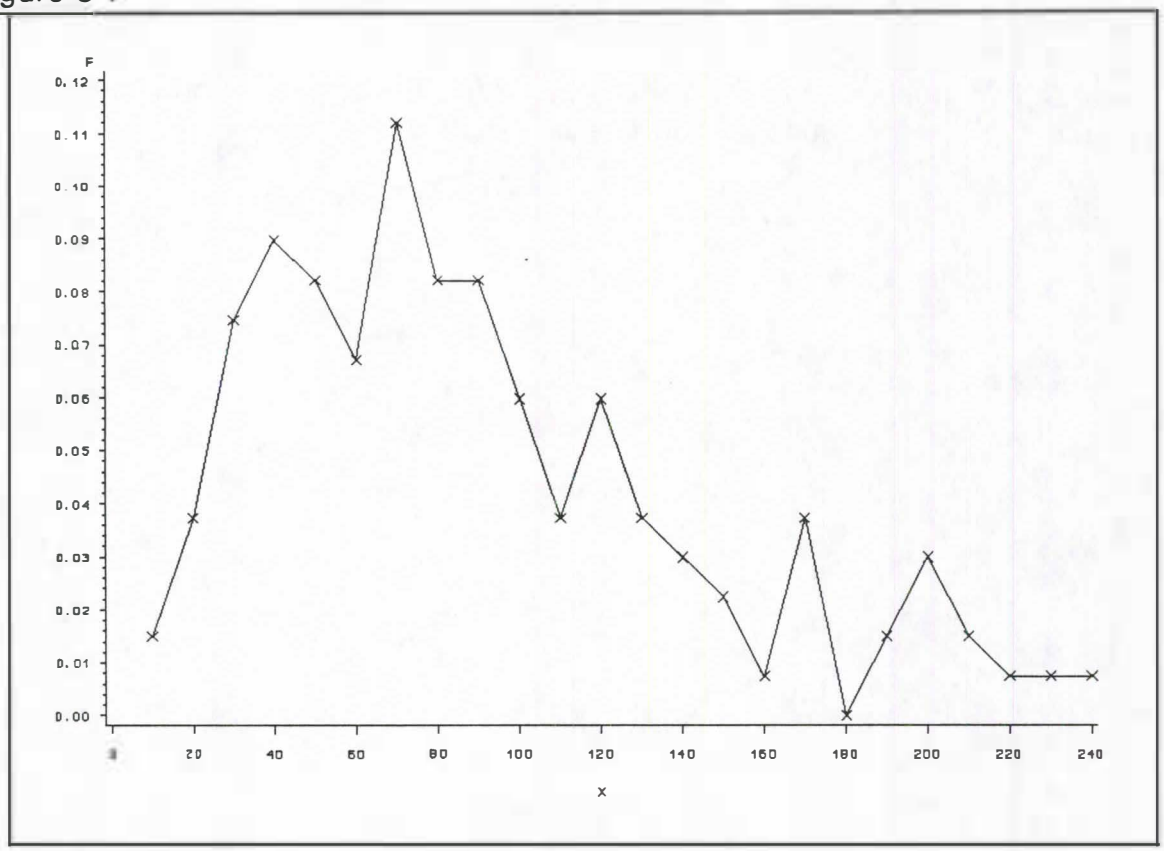
$$h = 2*3^{1/3}*\pi^{1/6}*\sigma*n^{-1/3} = 3{,}49*\sigma*n^{-1/3}$$

6. Les polygones de fréquence :

Les polygones de fréquence se construisent comme de simples graphes de points, dont l'abscisse est égale au milieu de la classe et l'ordonnée à la fréquence sur l'amplitude f_i/a_i. Puis ces points sont reliés par des droites.

SCOTT propose que pour la densité d'une loi normale de variance σ^2, la largeur de la classe soit égale à $h = 2,15*\sigma*n^{-1/5}$.

Figure 8 :



7. Un gramme de SAS et beaucoup d'histo :

Le logiciel statistique SAS® permet de dessiner des histogrammes, avec les procédures **CHART** pour les semi-graphiques (ou graphiques avec des caractères) (Figure 10) ou **GCHART** pour les graphiques (Figure 9).

La représentation par SAS est cependant plus proche des tuyaux d'orgues que de celle des histogrammes, à cause d'un espace inter-rectangle, qu'il est possible d'éliminer par l'option SPACE = 0 pour les graphiques (Figure 9 droite).

Le calcul du nombre de classes de l'histogramme est différent suivant le numéro de version du logiciel. En effet, la version 5 de SAS utilisait un algorithme de NELDER, tandis que la version 6 utilise l'algorithme de TERRELL et SCOTT.

Figure 9 :

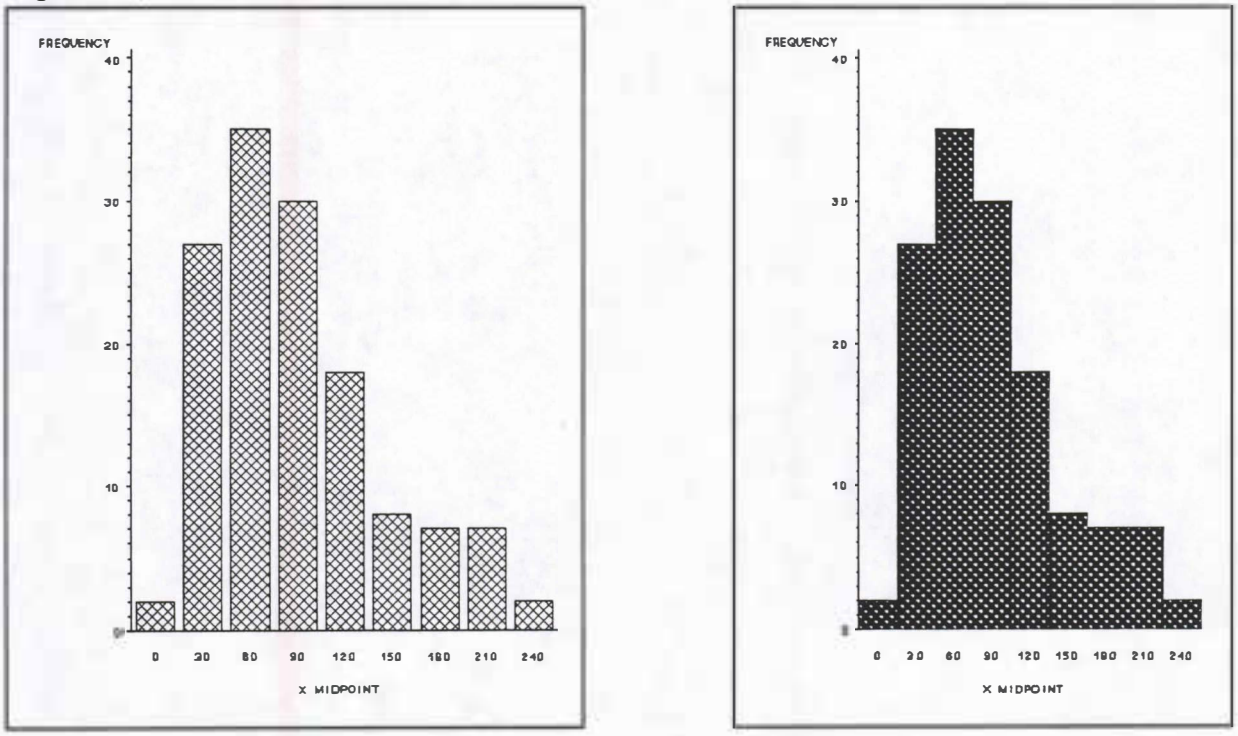
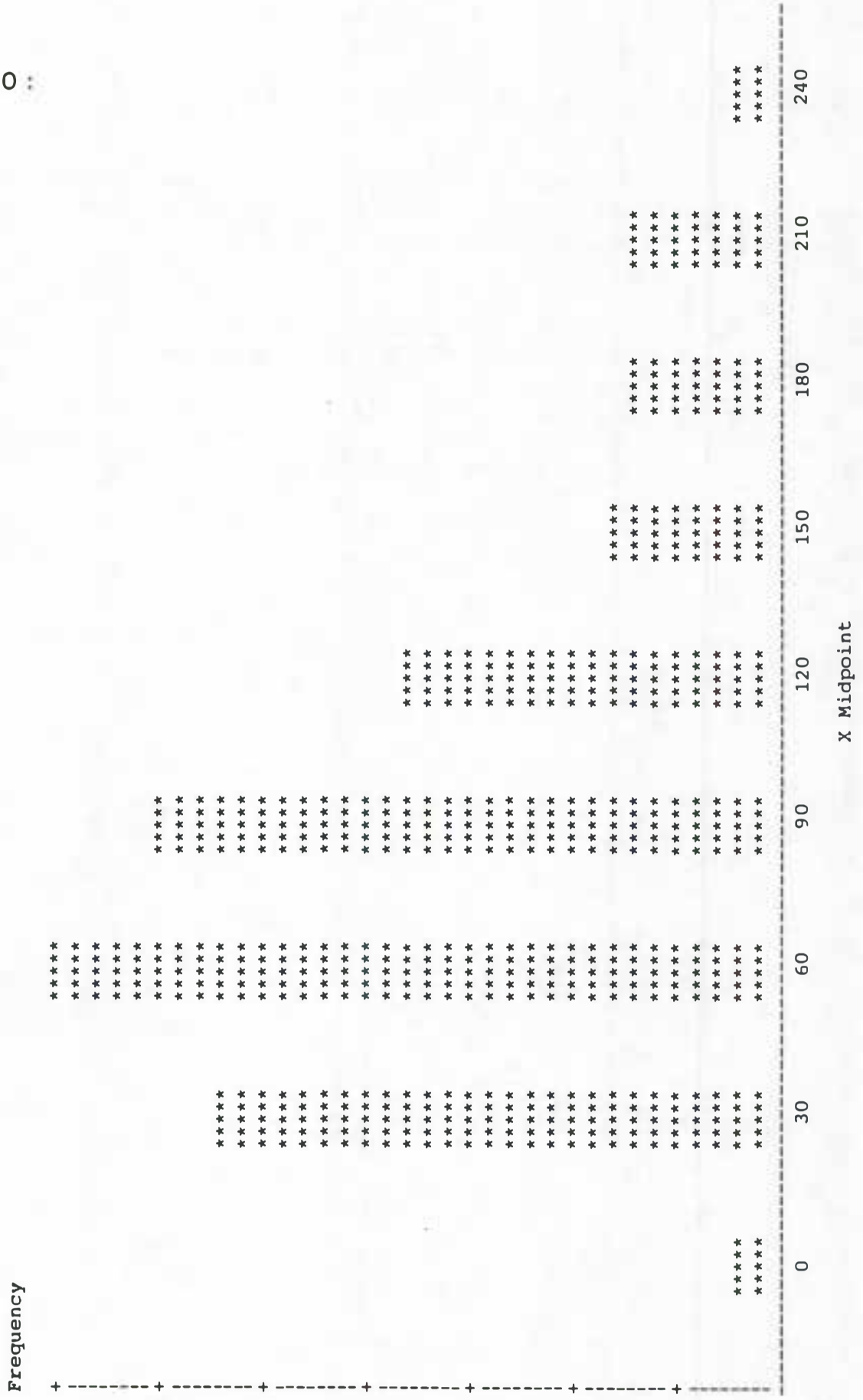


Figure 10 :


Frequency
35
30
25
20
15
10
5
0
30
60
90
120
150
180
210
240
X Midpoint

Références

CALOT Gérard (1973), "Cours de statistique descriptive", Deuxième édition, *Dunod*, 483 pages.

EMERSON John D. and HOAGLIN David C., "Stem-and-Leaf Displays", Understanding Robust and Exploratory Data Analysis, pages 7-30.

NELDER J.A. (1976), "A Simple Algorithm for Scaling Graphs", *Applied Statistics*, Volume 25, Numéro 1, LONDON : The Royal Statistical Society

SCOTT David W. (1979), "On optimal and data-based histograms", *Biometrika*, 1979, Volume 66, Numéro 3, pages 605-610.

SCOTT David W. (1985), "Frequency Polygons : Theory and Application", *Journal of the American Statistical Association*, June 1985, Volume 80, Numéro 390, pages 348-354.

TERRELL George R. and SCOTT David W. (1985), "Oversmoothed Nonparametric Density Estimates", *Journal of the American Statistical Association*, March 1985, Volume 80, Numéro 389, pages 209-214.

STEM AND LEAF
ou
TIGE ET FEUILLE

Septembre 1992

Hervé LEDOUX

BIOMETRIE
CIRAD - Forêt

Stem and Leaf
ou
Tige et Feuille

1. Introduction :

Le graphique "Stem and Leaf" que nous pouvons traduire par "Tige et Feuille" est utilisé pour représenter la distribution d'une variable. Ce type de graphique permet de visualiser ou de détecter :

. la symétrie de la distribution
. l'étendue et la dispersion
. si quelques valeurs sont éloignées du reste
. s'il existe des concentrations de valeurs
. ou au contraire s'il existe des trous

Le Stem and Leaf ressemble aux histogrammes, mais présente cependant quelques avantages :

. Il est plus facile à construire à la main.
. Il est plus facile de positionner la médiane ou d'autres statistiques basées sur l'ordre des données.
. Il est possible de distinguer la distribution des données à l'intérieur d'un intervalle.

Il n'existe pas de théorie élaborée concernant la construction du Stem and Leaf, car ce type de graphique nécessite une interactivité entre l'analyste et le graphique, en particulier dans le choix du nombre d'intervalles.

Nous commencerons par la présentation d'un Stem and Leaf simple, nous détaillerons les étapes de sa construction, puis nous présenterons quelques variations de ce graphique.

2. Construction du Stem and Leaf :

Pour expliquer la construction du Stem and Leaf, nous débuterons avec un exemple de 21 données, correspondant à la température maximale relevée dans certaines villes.

Les données, que nous avons triées par ordre croissant, sont les suivantes :

22,9	26,3	26,6	26,8	26,9	26,9	27,5
27,6	27,6	28,0	28,4	28,4	28,5	28,8
28,8	29,4	29,9	30,0	30,3	31,2	31,8

Pour débuter, il convient de choisir une paire de chiffres adjacents dans les données. Dans cet exemple, pour la donnée 22,9, nous pouvons choisir 22 ou 29. La paire de chiffres sera découpée et la partie gauche de la donnée constituera la tige tandis que le premier chiffre de la partie droite constituera la feuille. Dans l'exemple nous choisissons de séparer la partie entière de la partie décimale.

Donnée		Séparation		Stem and Leaf
22,9	⇒	22 | 9	⇒	22 et 9

Ensuite nous construisons une ligne verticale de séparation, de façon à pouvoir afficher toutes les tiges. Dans cet exemple, la ligne verticale doit pouvoir afficher 10 tiges (de 22 à 31). Puis nous affichons sur la droite de cette ligne les premiers chiffres, correspondant aux feuilles de chaque donnée.

Figure 1:

22	9
23	
24	
25	
26	36899
27	566
28	044588
29	49
30	03
31	28

Profondeur des données :

A chaque donnée, on peut assigner deux rangs, le premier rang correspondant au rang par ordre croissant, et le deuxième au rang par ordre décroissant. La "profondeur" d'une donnée correspond au plus petit de ces deux rangs. Ainsi la donnée 26,3 a un rang égal à 2 par ordre croissant et égal à 20 par ordre décroissant, sa profondeur est égale à 2. Sur le graphique, on inscrit dans la colonne de gauche, la plus grande profondeur correspondant aux feuilles d'une tige, excepté la tige contenant la médiane, qui reçoit le nombre de feuilles de la ligne, ce nombre est présenté entre parenthèses (Si le nombre de données est pair et que la médiane se situe entre deux lignes, nous ne représentons pas le nombre de feuilles). Dans notre exemple, la médiane est égale à 28,4, et la colonne profondeur de la tige 28 reçoit la valeur (6).

Cette technique permet, lorsque le graphique est construit à la main, de vérifier que l'on n'a pas omis de points, car le nombre de la ligne médiane plus les

profondeurs des deux lignes adjacentes est égal au nombre de données. Dans notre exemple, nous avons (6) + 9 + 6 = 21.

Figure 2:

Profondeur		
1	22	9
	23	
	24	
	25	
6	26	36899
9	27	566
(6)	28	044588
6	29	49
4	30	03
2	31	28

Nombre de lignes d'un Stem and Leaf :

Le choix du nombre de lignes d'un Stem and Leaf est important, car de celui-ci dépend le jugement que l'on peut porter sur le graphique. Ce nombre de lignes est fonction du nombre de données, et de l'étendue de celles-ci. Ce problème de détermination du nombre de lignes ou de l'intervalle des tiges est identique à celui rencontré pour la construction des histogrammes.

Un moyen simple de déterminer le nombre de lignes est le suivant :

Calculons $L = Int[10*\log_{10}(n)]$, où Int[] représente la partie entière et n le nombre de données. Cette valeur correspond au nombre maximum de lignes à afficher. Cette règle semble donner de bons résultats pour un nombre de données compris entre 20 et 300. Pour notre exemple, nous avons $L = Int[10*\log_{10}(21)] = 13$.

Calculons E l'étendue des données, dans notre exemple E = 31,8 - 22,9 = 8,9. Puis calculons le rapport de E sur L, qui représente l'intervalle d'une ligne. Nous arrondissons à la plus proche puissance de 10 par valeur supérieure. Ainsi nous avons la largeur de l'intervalle des lignes. Dans notre exemple E/L = 0,68, que nous arrondissons à $1 = 10^0$. Nous avons utilisé cette valeur pour dessiner les figures 1 et 2.

3. Variation sur le même Stem :

Chaque tige d'un Stem and Leaf reçoit les feuilles comprises entre 0 et 9. Mais dans certains cas, le graphique peut être trop tassé. Une solution est de découper les tiges en deux (Cf. Figure 3), la tige 0 se transforme en tiges 0* et 0. qui reçoivent respectivement les feuilles comprises entre 0 - 4 et 5 - 9.

Figure 3 :

Cette variante est illustrée par un exemple représentant la dureté de l'aluminium :

53,0	70,2	84,3	55,3	78,5	63,5	71,4
53,4	82,5	67,3	69,5	73,0	55,7	85,8
95,4	51,1	74,4	54,1	77,8	52,4	69,1
53,5	64,3	82,7	55,7	70,5	87,5	50,7
72,3	59,5					

Figure 4 :

11	5	01233345559
(5)	6	34799
14	7	00123478
6	8	22457
1	9	5

7	5*	0123334
11	5.	5559
13	6*	34
(3)	6.	799
14	7*	001234
8	7.	78
6	8*	224
3	8.	57
	9*	
1	9.	5

En appliquant la régle de détermination du nombre de tiges, nous trouvons L = $Int[10*\log_{10}(30)] = 14$, E = 95,4 - 50,7 = 44,7 et E/L = 3,19. En arrondissant cette valeur, nous trouvons que l'intervalle d'une tige est de 10 (Figure 4 gauche), alors que la valeur est plus proche de 5 (Figure 4 droite).

Au lieu de découper la tige en deux, il est aussi possible de la découper en 5. La tige 0 se transforme en 0*, t (two), f(four), s(six), 0., qui contiennent respectivement les feuilles égales à 0-1, 2-3, 4-5, 6-7, 8-9 (Figure 5).

Figure 5 :

Une autre variante, est d'inscrire tous les chiffres à droite de la ligne de séparation, et de séparer les feuilles par des virgules.

La figure 6 représente à gauche un Stem and Leaf classique et à droite la variante. Les données utilisées dans cet exemple sont le prix de 17 Chevrolet en dollars :

150	250	688	695	795	795	895
895	895	1099	1166	1333	1499	1693
1699	1775	1895				

Figure 6 :

1	5
2	5
3	
4	
5	
6	89
7	99
8	999
9	
10	9
11	6
12	
13	3
14	9
15	
16	99
17	7
18	9

1	50
2	50
3	
4	
5	
6	88,95
7	95,95
8	95,95,95
9	
10	99
11	66
12	
13	33
14	99
15	
16	93,99
17	75
18	95

4. SAS et le Stem and Leaf :

Pour dessiner un Stem and Leaf avec le logiciel statistique SAS®, il faut employer la procédure **UNIVARIATE**.

Cette procédure imprime un certain nombre de statistiques relatives à la variable étudiée (moyenne, moments, quantiles, extrêmes, ...), ainsi que des graphiques (semi-graphique ou graphique avec des caractères), tel que le Stem and Leaf, le Box-Plot ou le Normal Probability Plot.

Le Stem and Leaf affiché par SAS ne contient pas la colonne "Profondeur", mais une colonne (#) correspondant à l'effectif de la tige. L'exemple du paragraphe 2, est représenté dans la figure 7.

Figure 7 :

Stem	Leaf	#
31	28	2
30	03	2
29	49	2
28	044588	6
27	566	3
26	36899	5
25		
24		
23		
22	9	1

BOX PLOTS

Septembre 1992

Matthieu LESNOFF

BIOMETRIE
CIRAD - Forêt

BOX PLOTS

Le *box plot* est une méthode simple développée par TUKEY (1977), qui permet de résumer la structure d'un ensemble de données. Par une impression visuelle rapide, le *box plot* renseigne sur les traits caractéristiques de la distribution des données observées (distribution empirique) :

- paramètre central (localisation de la distribution)
- paramètre de dispersion
- étalement de la distribution
- existence de valeurs extrêmes
- symétrie / asymétrie.

Cette méthode est particulièrement utile pour comparer plusieurs ensembles de données (juxtaposition de *box plots*), lorsqu'il n'est pas nécessaire ou impossible de détailler chacune des distributions.

I - PRINCIPE DU BOX PLOT

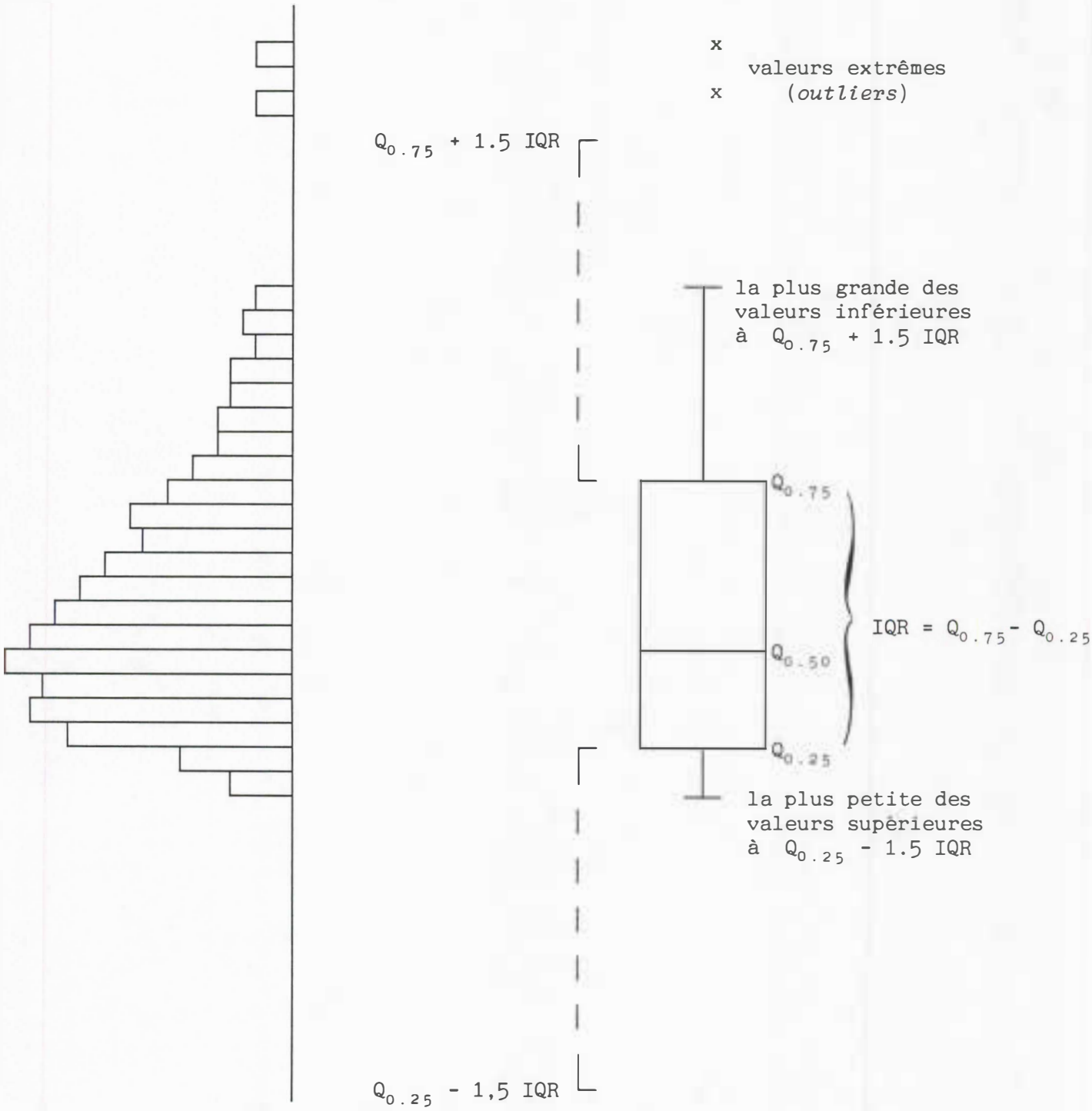
1. Construction du box plot

La construction du *box plot* est basée sur la détermination de quantiles empiriques particuliers : la médiane $Q_{0.50}$, le premier quartile $Q_{0.25}$ et le deuxième quartile $Q_{0.75}$. D'après les deux quartiles, on définit l'intervalle interquartile (*interquartile range*) :

$$IQR = Q_{0.75} - Q_{0.25}$$

IQR est une mesure de dispersion de la distribution.

.../...



Les premier et deuxième quartiles sont représentés par les parties inférieure et supérieure du rectangle (*box*). La ligne horizontale intermédiaire représente la médiane. En général, la largeur du *box* n'a pas de signification, sauf dans certaines variantes du *box plot* (cf. **III - 1.**).

.../...

D'après la définition originale de TUKEY (1977), les bras du *box plot* sont construits de la manière suivante :

- le bras supérieur s'étend jusqu'à la plus grande valeur inférieure ou égale à $Q_{0.75}$ + 1.5 IQR ;

- le bras inférieur s'étend jusqu'à la plus petite valeur supérieure ou égale à $Q_{0.25}$ - 1.5 IQR .

Toute donnée observée en dehors de l'intervalle $I = [Q_{0.25} - 1.5\ IQR\ ,\ Q_{0.75} + 1.5\ IQR]$ est représentée par un point individuel et est considérée comme une valeur extrême ou "donnée aberrante" (*outlier*). Une valeur extrême n'est pas forcément aberrante (dans le sens de valeur erronée) ; elle peut simplement appartenir à la queue "normale" de la distribution.

Par des simulations par rapport à une loi normale, HOAGLIN et al. (1986) ont montré que la valeur du facteur multiplicatif de l'IQR ($k = 1.5$) , qui détermine l'étendue de l'intervalle I , convenait bien pour l'identification de valeurs extrêmes méritant une attention particulière. Cette définition correspond à un compromis entre un nombre raisonnable de données extrêmes et un faible risque d'en laisser échapper une.

Une propriété importante du *box plot* est sa résistance aux perturbations pouvant affecter les données (valeurs extrêmes, erreurs de mesures, etc...). A la différence de la moyenne et de l'écart-type, la médiane et les quartiles sont en effet peu sensibles à l'existence de données extrêmes dans un échantillon. Jusqu'à 25 % des valeurs d'un ensemble de données peuvent ainsi être transformées en valeurs très fortes sans perturber la forme générale du *box plot* (HOAGLIN et al., 1983).

Remarques :

- Certains auteurs se sont libérés de la définition de TUKEY pour la construction des *box plots*. Par exemple, IGLEWICZ et HOAGLIN (1987) utilisent une moyenne α-tronquée ($\alpha = 25\ \%$) comme centre de distribution à la place de la médiane. CLEVELAND (1985) remplace l'intervalle I par l'intervalle $[Q_{0.10} , Q_{0.90}]$ pour détecter les valeurs extrêmes.

- Plusieurs paramètres centraux peuvent être indiqués sur le *box plot* (exemple : la moyenne peut être ajoutée à la médiane dans le rectangle).

- Il reste le problème du calcul des quantiles empiriques de la distribution. Différentes méthodes d'estimation sont proposées et, à l'heure actuelle, il n'y a pas de standardisation, en particulier au niveau des logiciels statistiques. Ces méthodes peuvent influencer assez fortement la forme générale du *box plot*, surtout pour les échantillons de petite taille.

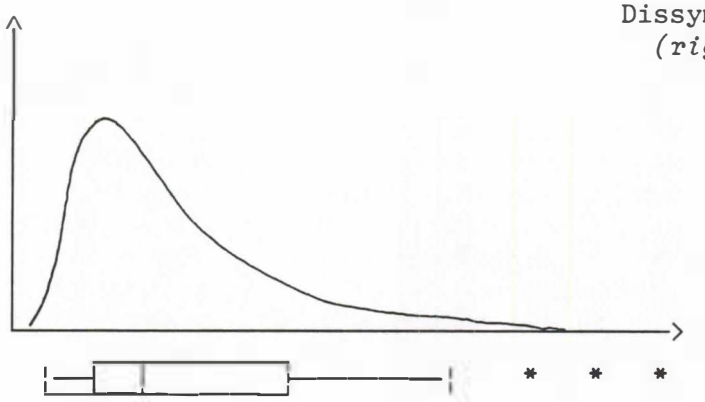
.../...

2. Box plot d'une distribution empirique

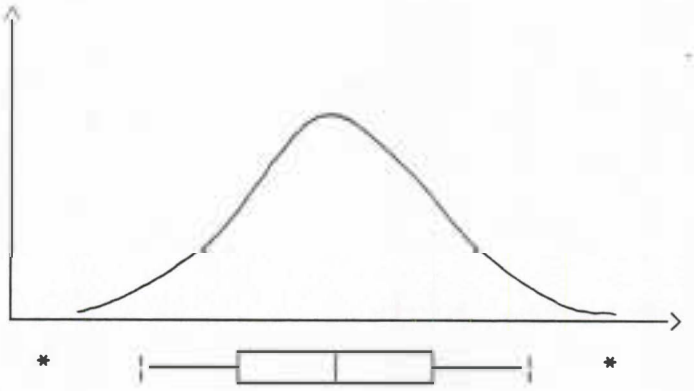
a) Traits caractéristiques :

Le centre et la dispersion de la distribution sont représentés par la médiane et l'intervalle IQR. Plus la taille du rectangle est grande, plus la dispersion est grande. La position centrée ou excentrée de la médiane dans le rectangle renseigne sur la symétrie ou l'asymétrie de la distribution.

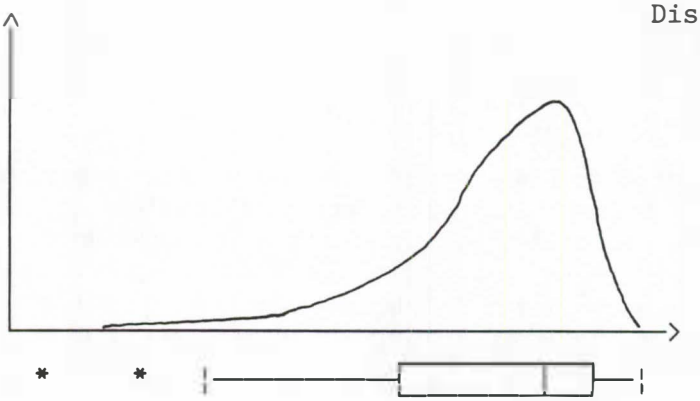
Dissymétrie à droite
(right skewness)



Symétrie

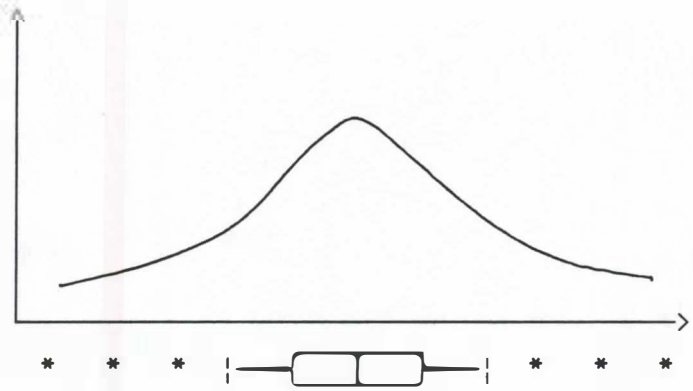


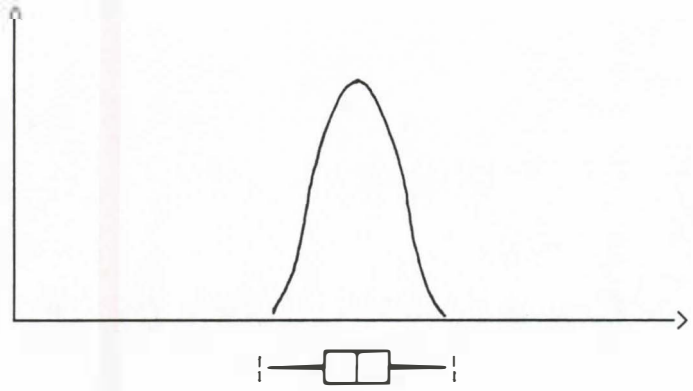
Dissymétrie à gauche
(left skewness)



.../...

Plus la distribution aura des queues lourdes, plus il y aura des valeurs extrêmes sur le *box plot*.

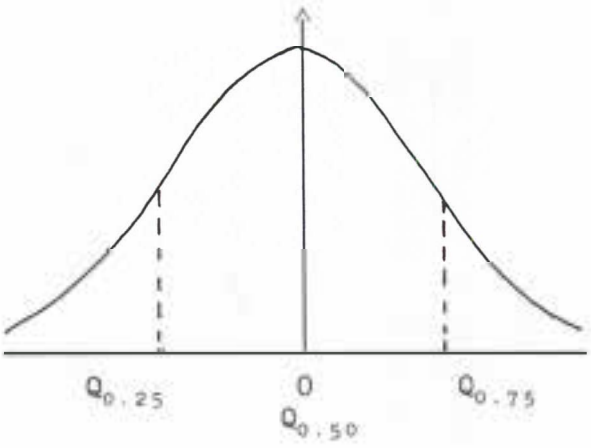




.../...

b) Box plots de lois classiques

Box plot d'une distribution normale $N(0.1)$:



$Q_{0.25} = -0.67$

$Q_{0.50} = 0$

$Q_{0.75} = 0.67$ IQR = 1.35

soit X : $N(0.1)$

$P(-2.7 < X < 2.7) = 0.993$

$-0.67 - 1.5$ IQR $= -2.7$ $\qquad$ $0.67 + 1.5$ IQR $= 2.7$

==> le *box plot* d'une distribution $N(0.1)$ contient 99.3 % de la population. 0.7 % de la population sont des valeurs extrêmes.

Le *Tableau n° 1* donne les pourcentages de valeurs extrêmes (*% out*) pour d'autres lois (uniforme, Student, Chi-deux). Plus les distributions sont étalées et ont des queues lourdes, plus ce pourcentage augmente. Il est nul pour une distribution uniforme sur [- 1.1] .

.../...

Distribution	M^a	Upper[b] Fourth	Outlier[c] Cutoffs	Total[d] % Out	Value[e] of 1.96σ	% Outside $\mu \pm 1.96\sigma$
			Symmetric			
$U(-1, 1)$	0	0.500	±2.000	none	1.132	none
$N(0, 1)$	0	0.674	±2.698	0.70	1.960	5.00
t_{20}	0	0.687	±2.748	1.24	2.066	5.20
t_{10}	0	0.700	±2.800	1.88	2.191	5.32
t_5	0	0.727	±2.908	3.35	2.530	5.25
t_1	0	1.000	±4.000	15.59	—	—
			Nonsymmetric			
χ^2_1	0.45	0.102	−1.730[f]	7.58	−1.772	5.22
		1.323	3.155		3.772	
χ^2_5	4.35	2.675	−3.252[f]	2.80	−1.198	4.78
		6.626	12.552		11.198	
χ^2_{20}	19.34	15.452	2.888	1.39	7.604	4.53
		23.828	36.392		32.396	

[a] M = median of distribution. Defined so that $F(M) = .5$, where F is the cumulative distribution function.
[b] Upper fourth is the value above which .25 of the probability lies. (Lower fourth has .25 of probability below it.) For the nonsymmetric distributions, the entries in this column are the lower fourth and the upper fourth.
[c] Upper outlier cutoff = upper fourth + $\frac{3}{2}$ × (upper fourth − lower fourth). (Lower outlier cutoff = lower fourth − same quantity.)
[d] % Out = percent of probability below the lower outlier cutoff or above the upper outlier cutoff.
[e] For $U(-1, 1)$, $\sigma = \sqrt{1/3} \approx .58$ and $\mu = 0$. For t_ν, $\sigma = \sqrt{\nu/(\nu - 2)}$ for $\nu > 2$ and $\mu = 0$. For χ^2_ν, $\mu = \nu$ and $\sigma = \sqrt{2\nu}$.
[f] For skewed distributions, one of the pair of cutoffs often falls beyond the range of possible values.

Tableau n° 1 : Valeurs des médianes, quartiles et pourcentage de valeurs extrêmes pour différentes distributions (d'après HOAGLIN et al., 1983)

Si l'on travaille sur un échantillon suffisamment grand, on doit retrouver les pourcentages théoriques de valeurs extrêmes du *Tableau n° 1*. Ceci permet, par exemple, en observant le *box plot* d'une distribution empirique, de juger si ses queues sont plus (ou moins) lourdes qu'une distribution $N(0.1)$ (le pourcentage de valeurs extrêmes dépasse-t-il 0.7 % ?).

.../...

Par contre, le *box plot* est beaucoup plus instable pour les échantillons de faibles tailles et le pourcentage de valeurs extrêmes d'une distribution théorique deviendra très variable (exemple : le *box plot* d'un échantillon de 10 individus issus d'une loi $N(0.1)$ n'aura pas un pourcentage de valeurs extrêmes égal à 0.7 %).

Cette instabilité souligne le fait que, si l'on veut comparer des distributions empiriques à l'aide de *box plots*, il faut chercher à comparer des échantillons de tailles semblables (un *box plot* construit sur un échantillon de 10 individus n'est pas de même nature que celui construit sur un échantillon de 1.000 individus).

II - BOX PLOTS ET COMPARAISONS DE DISTRIBUTIONS

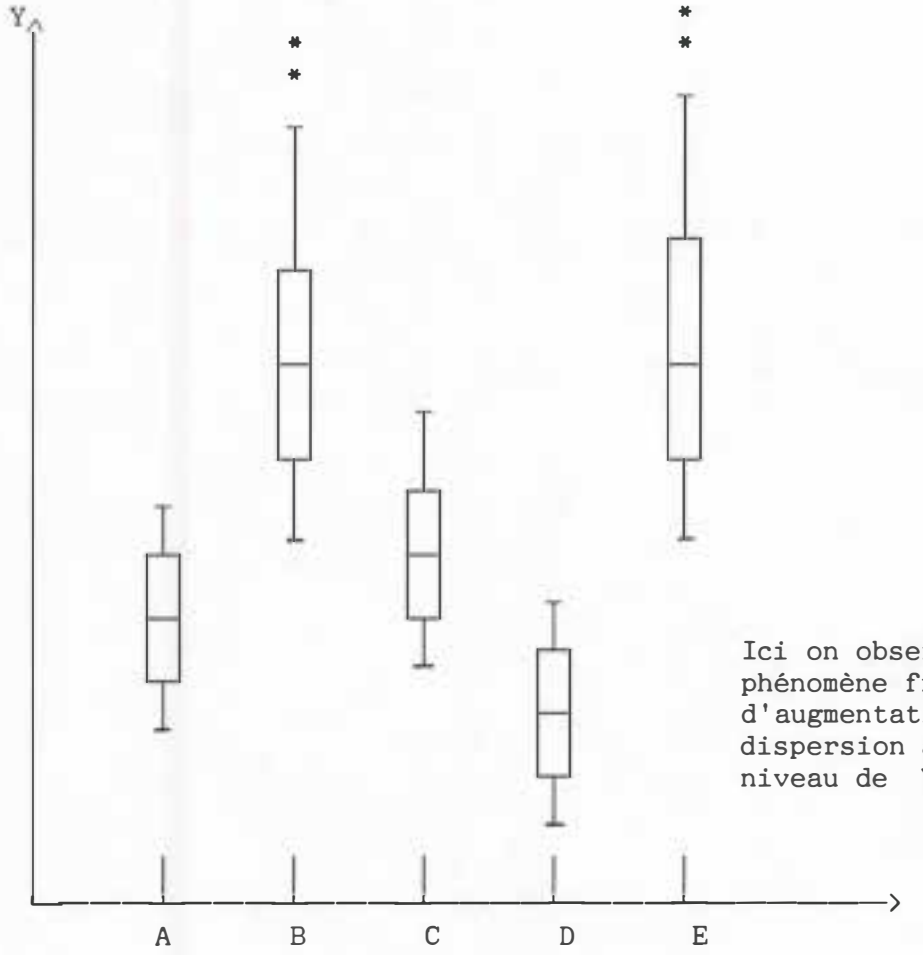
La juxtaposition de *box plots* sur un même graphique facilite la comparaison de plusieurs ensembles de données. Celle-ci s'effectue en fonction des traits caractéristiques de chaque distribution dégagés par les *box plots* :

- localisation → position relative des médianes
- dispersion → tailles des rectangles
- étalement → longueurs des bras
- symétrie → positions des médianes dans les rectangles
- valeurs extrêmes → points individuels.

.../...

1. Box plots en fonction d'une variable nominale ou ordinale

Exemple : Facteur à 5 modalités (A, B, C, D, E)

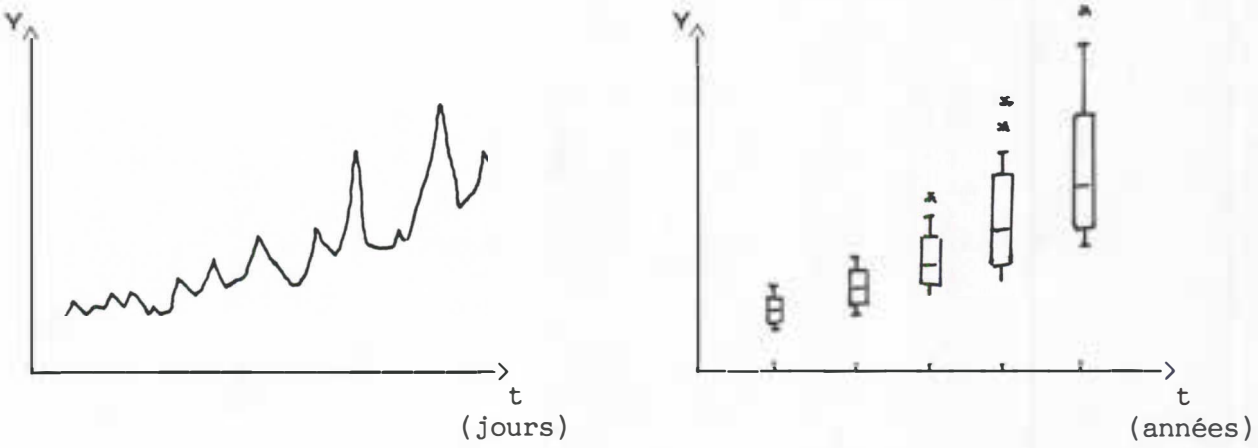


Ici on observe le phénomène fréquent d'augmentation de la dispersion avec le niveau de Y .

.../...

2. Box plots et série chronologiques

La technique du *box plot* permet d'étudier graphiquement la stationnarité à l'ordre 2 d'une série temporelle : Existe-t-il une tendance ? La variance est-elle constante ?



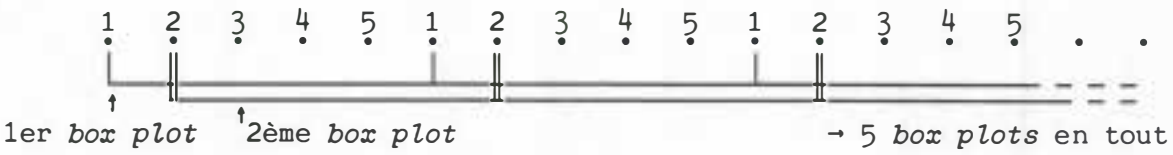
(chaque *box plot* regroupe 365 observations)

On observe nettement sur ces graphiques une tendance (augmentation du niveau de la série) et une variance instable liée au niveau de X . La série n'est pas stationnaire.

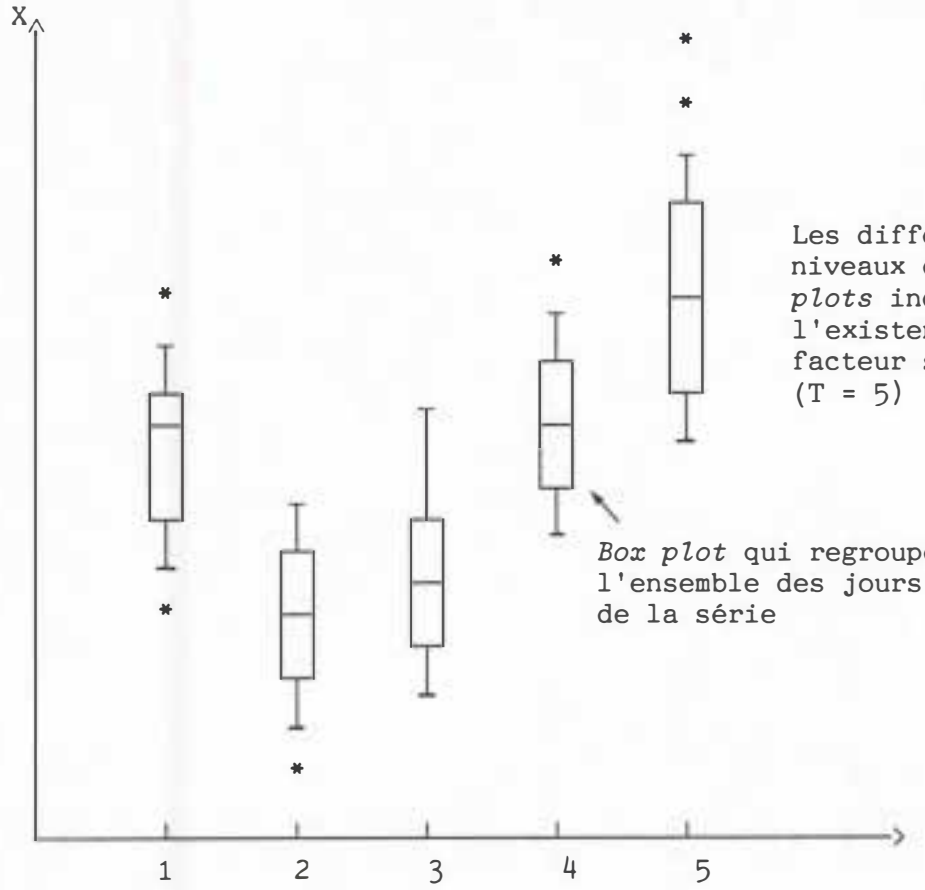
En outre, le *box plot* est très utile pour détecter un facteur saisonnier :

On choisit un pas de temps T (ex. : T = 5 j.) et on regroupe dans chaque *box plot* toutes les observations homologues par rapport à T sur l'ensemble de la série.

Exemple : T = 5 j

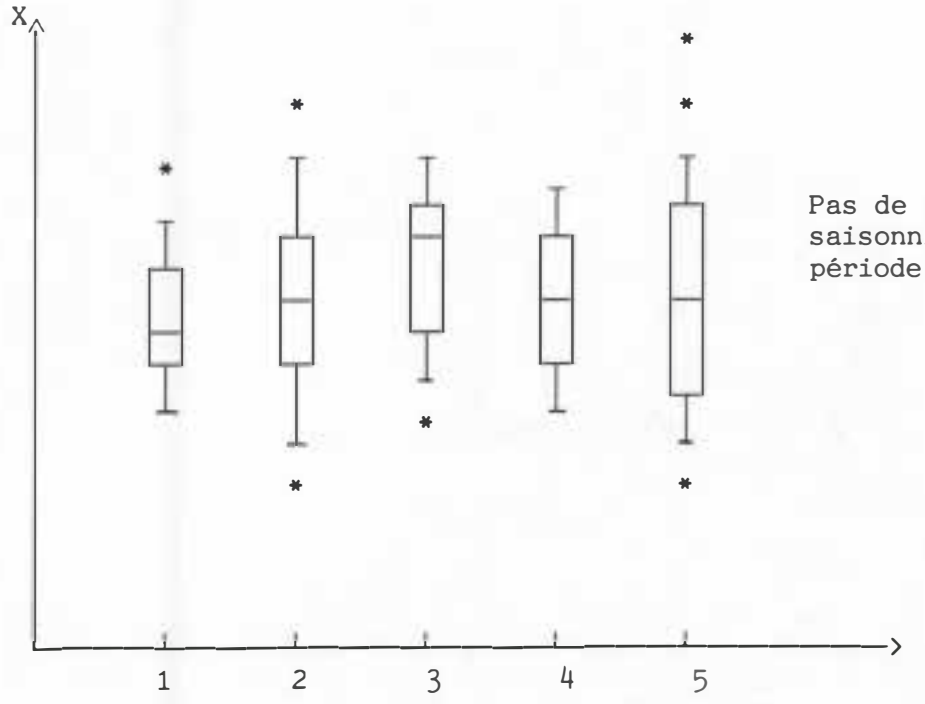


.../...



Les différences de niveaux des *box plots* indiquent l'existence d'un facteur saisonnier (T = 5)

Box plot qui regroupe l'ensemble des jours "4" de la série

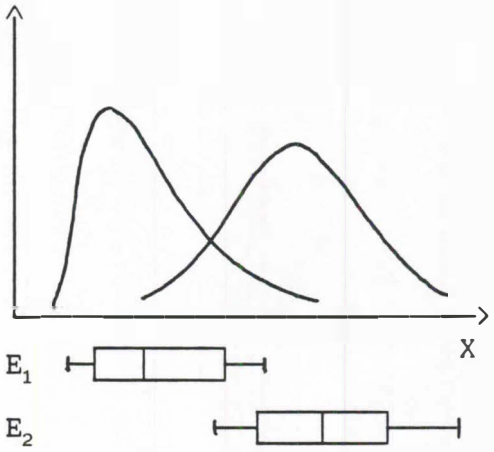


Pas de facteur saisonnier de période 5 visible.

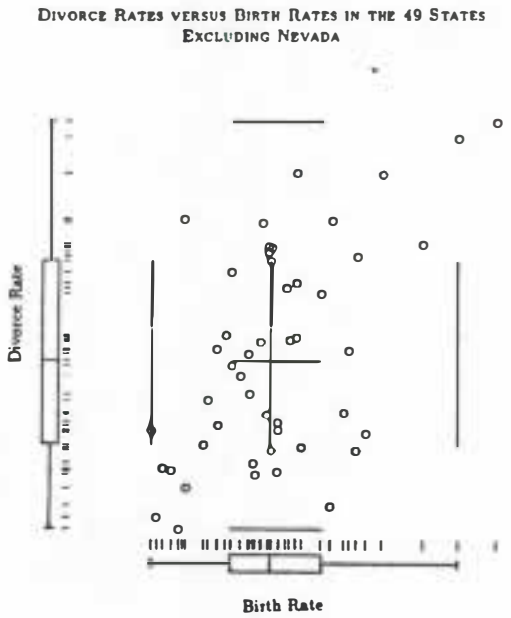
.../...

3. Autres types de graphiques avec box plots

* Distributions de fréquences avec *box plots*



* Graphes (X, Y) avec *box plots* marginaux et *rangefinder box plot*



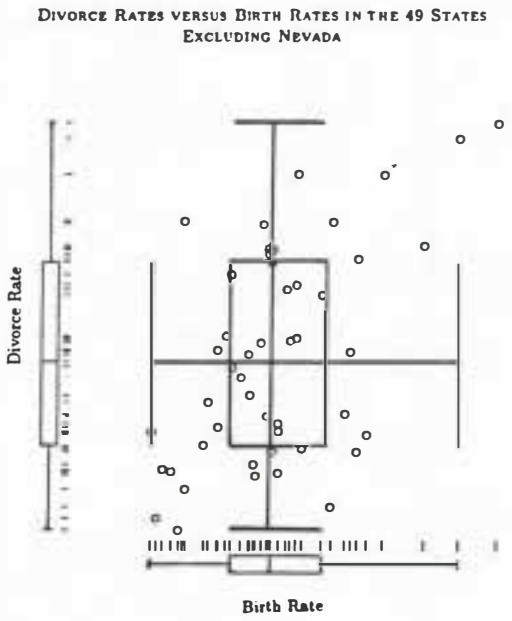
Les *box plots* représentés le long des axes du graphe de dispersion renseignent sur les distributions marginales de X et Y . En plus de ces *box plots* marginaux, BECKETTI et GOULD (1987) ont surimposé au nuage un *rangefinder box plot*.

.../...

Le *rangefinder box plot* contient exactement la même information que les *box plots* des variables X et Y . Les lignes centrales verticales et horizontales représentent respectivement l'interquartile de la distribution de Y et l'interquartile de la distribution de X . Ces deux lignes forment une croix qui constitue le corps du *rangefinder box plot*. Elles sont centrées respectivement sur les médianes de X et de Y . Les lignes verticales et horizontales excentrées représentent les limites des bras des deux *box plots* marginaux. Ce type de graphique permet en particulier de repérer directement les valeurs extrêmes ("points aberrants") pour l'une ou les deux variables.

On remarque que les axes (X,Y) du graphique n'ont pas été tracés. Ce choix va dans le sens de la théorie de TUFTE (1983) qui cherche à réduire le plus possible la quantité de signaux visuels non informatifs sur le graphique.

Il est possible d'étendre l'idée du *rangefinder box plot* en construisant directement un *box plot* à deux dimensions : il suffit de compléter le graphique précédent en traçant le corps et les bras des *box plots*.

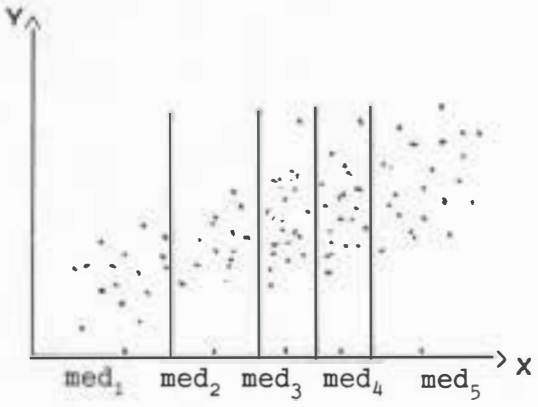


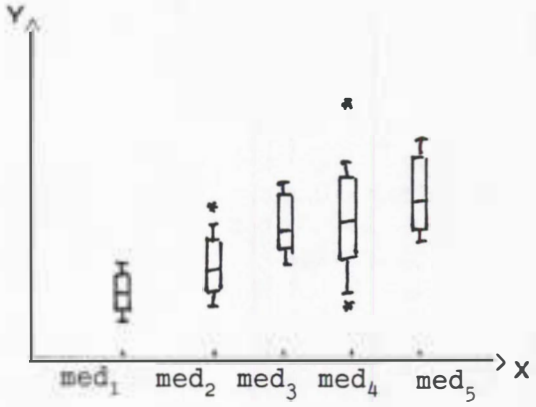
Le *box plot* à deux dimensions peut être très utile quand on veut comparer plusieurs nuages de points sur un même graphique. Chaque nuage peut être résumé par le corps d'un *box plot* à deux dimensions ; les bras peuvent être négligés pour augmenter la lisibilité du graphique.

.../...

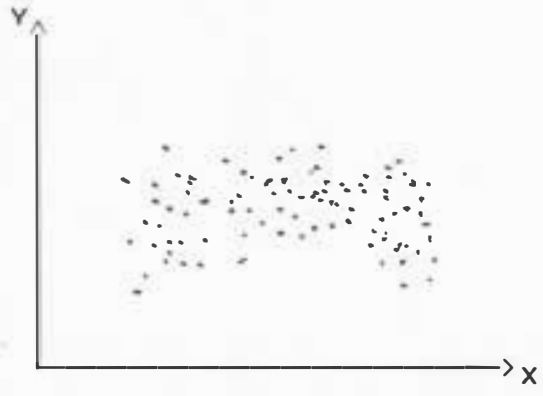
* Etude de la dépendance entre X et Y à l'aide de *box plots* :

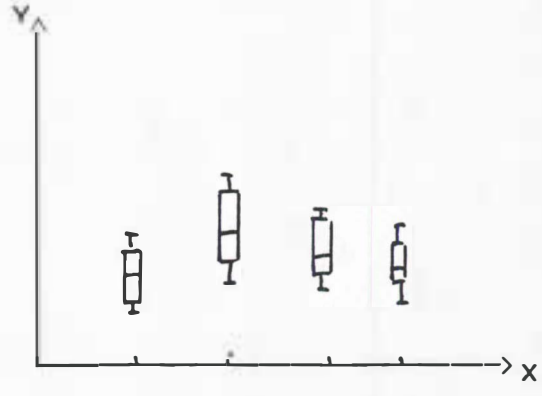
Pour étudier les variations de la distribution locale de Y en fonction de X , il est possible de diviser l'échantillon en n classes d'effectifs semblables par rapport à la variable X , puis de construire un *box plot* par classe (CLEVELAND, 1985 ; HOAGLIN et al., 1983). Au niveau de chaque classe i , le *box plot* est centré sur la médiane $med_i(X)$ de cette classe.





Ce type de graphe permet de mettre en évidence des structures qui n'apparaissent pas clairement sur un graphe (X, Y) classique :





Mise en évidence d'une forme en cloche peu visible sur le graphe (X,Y)

.../...

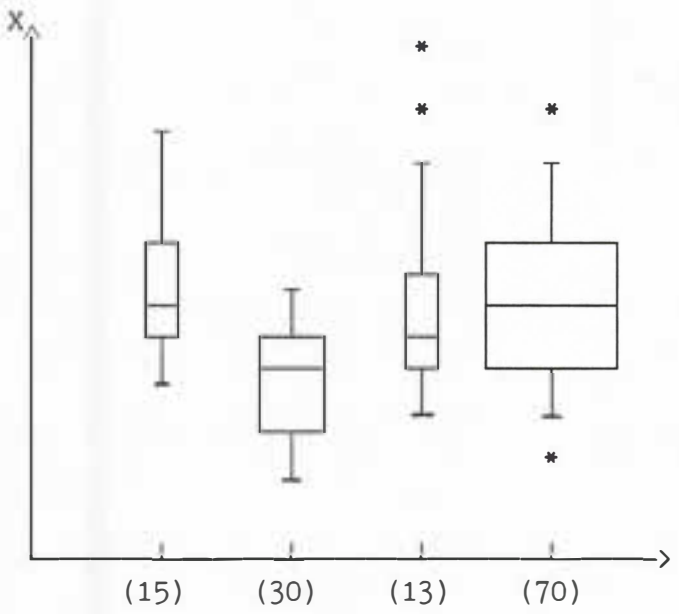
III - VARIANTES DU BOX PLOT

1. Box plot à largeurs variables

Il est important de connaître l'effectif des échantillons sur lesquels les *box plots* ont été construits, en particulier lorsqu'on veut comparer plusieurs ensembles de données (on ne compare pas de la même manière deux échantillons de tailles semblables et deux échantillons de tailles très différentes).

Une méthode est d'inscrire l'effectif de l'échantillon en dessous de chaque *box plot*. Une autre possibilité, proposée par McGILL et al. (1978), est de construire des *box plots* de largeurs variables : la largeur de chaque rectangle est proportionnelle à la racine carrée de la taille de l'échantillon correspondant.

L'objectif est de pouvoir apprécier directement, de manière visuelle, les différences d'effectifs des échantillons comparés, en attribuant des surfaces variables aux différents *box plots*.



Effectifs des échantillons

McGILL et al. justifient le choix de la racine carrée de l'effectif pour déterminer la largeur des *box plots* par le fait que de nombreuses mesures de dispersions, comme l'erreur standard, sont inversement proportionnelles à cette valeur.

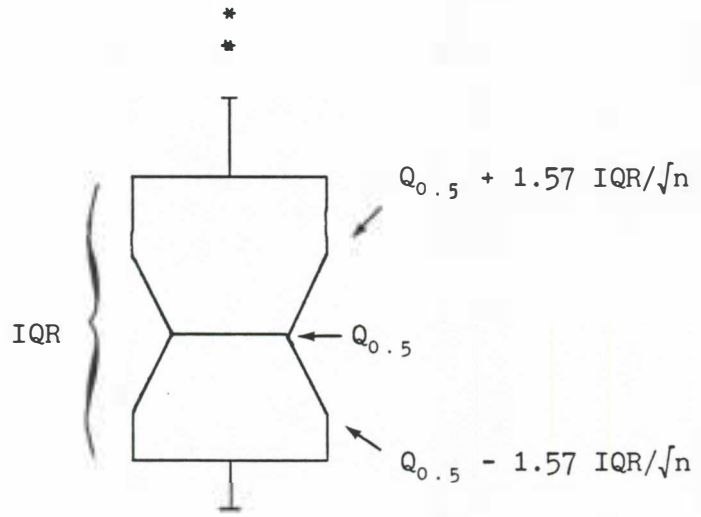
.../...

2. Box plots à encoches : "Notched Box Plots"

Les *box plots* à encoches (McGILL et al., 1978) permettent de "tester" de manière visuelle les différences de niveaux entre plusieurs distributions étudiées.

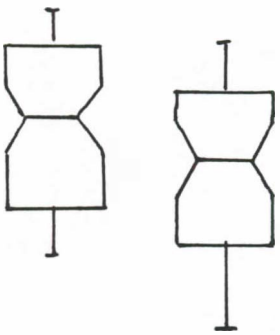
Le rectangle de chaque *box plot* est entaillé d'une encoche centrée sur la médiane et de taille :

$[Q_{0.5} \pm 1.57\ IQR/\sqrt{n}]$ (n étant l'effectif de l'échantillon) :

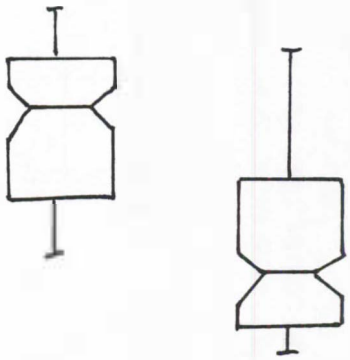


Lors d'une comparaison entre deux distributions, la différence entre les médianes est jugée significative si les encoches des *box plots* ne se chevauchent pas.

Différence non significative



Différence significative



.../...

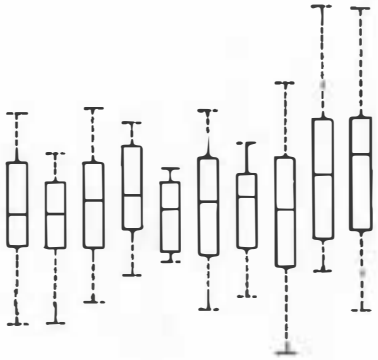
Qu'entend-on par "significatif" ? :

McGILL et al. (1978) ont montré que, sous l'hypothèse que les deux échantillons comparés soient indépendants et identiquement distribués et issus de deux populations de médianes inconnues ayant une forme normale au moins dans leur centre, la méthode des encoches (calculées avec la valeur $1.57\ IQR/\sqrt{n}$) fournit un test de l'hypothèse "H_0 : égalité des deux médianes" avec un risque approximatif de 5 % . Si les hypothèses ne sont pas strictement respectées (ce qui est fréquemment le cas), cette méthode reste très utile pour comparer la valeur des médianes.

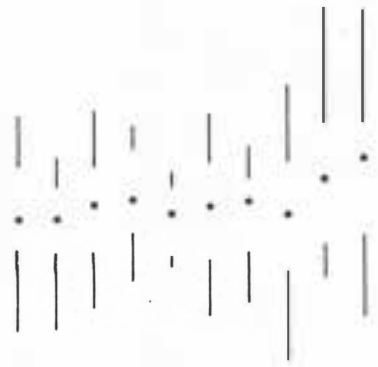
3. Autres variantes

Certains auteurs comme TUFTE (1983) ont cherché à simplifier au maximum les représentations graphiques en ne conservant qu'un minimum de signaux visuels informatifs : tout élément superflu est éliminé. Dans cet esprit, TUFTE a proposé une représentation schématique similaire aux *box plots* : les "*parallel schematic plots*". En conservant toute l'information des *box plots* de TUKEY, les schémas de TUFTE présentés ci-dessous permettent de réduire fortement le nombre d'éléments graphiques (barres verticales et horizontales). On peut remarquer qu'il est facile de suivre l'évolution de la médiane sur cette représentation, ce qui est beaucoup moins évident pour les *box plots* classiques.

Box plots
de TUKEY



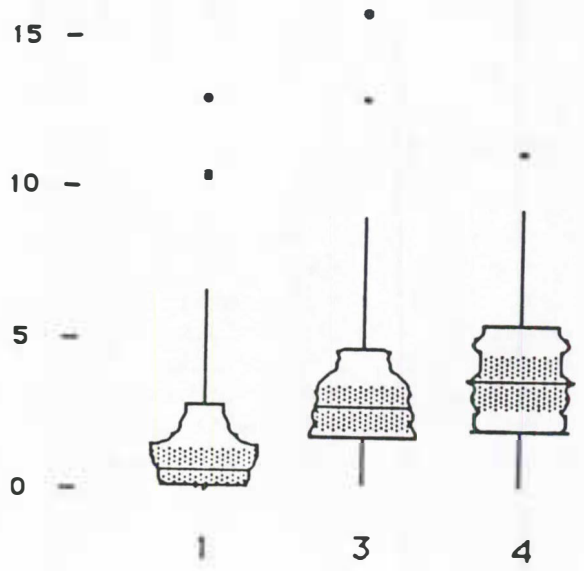
Parallel schematic plot
de TUFTE



.../...

D'autres auteurs ont, au contraire, cherché à ajouter de l'information aux *box plots* classiques de TUKEY. Par exemple, BENJAMINI (1988) propose le *vaseplot* : le *vaseplot* est un *box plot* dont la largeur en un point est proportionnelle à la densité locale estimée en ce point.

Vaseplot de BENJAMINI (Trois distributions du Chi-deux à 1, 3 et 4 degrés de liberté)



* * *

BIBLIOGRAPHIE

- **BECKETTI** (S.), **GOULD** (W.), 1987 - Rangefinder box plots - The American Statistician, 41(2) : 149-149.

- **BENJAMINI** (Y.), 1988 - Opening the box of a box plot - The American Statistician, 42(4) : 257-262.

- **CLEVELAND** (W.S.), 1985 - The element of graphical data - Wadsworth Advanced Books and Software, 323 p.

- **FRIGGE** (M.), **HOAGLIN** (D.C.), **IGLEWICZ** (B.), 1989 - Some implementation of the box plot - The American Statistician, 43(1) : 50-54.

- **HOAGLIN** (D.C.), **IGLEWICZ** (B.), **TUKEY** (J.W.), 1986 - Performance of same resistant rules for outlier labelling - JASA, 81 : 991-999.

- **HOAGLIN** (D.C.), **MOSTELLER** (F.), **TUKEY** (J.W.), 1983 - Understanding robust and exploratory data analysis - Wiley & Son, Inc., 447 p.

- **IGLEWICZ** (B.), **HOAGLIN** (D.C.), 1987 - Use of box plots for process evaluation - Journal of Quality Technology, 19 : 180-190.

- **McGILL** (R.), **TUKEY** (J.W.), **LARSEN** (W.A.), 1978 - Variations of box plots - The American Statistician, 32 : 12-16.

- **TUFTE** (E.), 1983 - The visual display of quantitative information - Cheshire, Connecticut : Graphic Press, 197 p.

- **TUKEY** (S.W.), 1977 - Exploratory data analysis - Addison Wesley, 489 p.

* * *

LE PERE QUANTILE

Septembre 1992

Jean-Claude BERGONZINI

BIOMETRIE
CIRAD - Forêt

BIOM/JCB/AG

LE PERE QUANTILE

Notations : Etant donné une série statistique $(x_1, x_2, \ldots x_k, \ldots x_n)$,

on notera :

$$x_{1:n} \quad x_{2:n} \quad x_{3:n} \ldots \quad x_{k:n} \ldots \quad x_{n:n}$$

la série ordonnée associée.

Si a est un nombre réel [a] désignera sa partie entière, X étant une variable aléatoire, F(x) désignera sa fonction de répartition et f(x) sa densité.

Exemple : $x_1 = 7$ $x_2 = 4$ $x_3 = 5$ $x_4 = 7$ $x_5 = 3$

$x_{1:5} = 3$ $x_{2:5} = 4$ $x_{3:5} = 5$ $x_{4:5} = 7$ $x_{5:5} = 7$

$a = 2{,}371$ $[a] = 2$ $a = -7{,}21$ $[a] = -7$

X de loi uniforme sur [0,1]

$$F(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ x & \text{si } 0 < x \leq 1 \\ 1 & \text{si } 1 < x \end{cases} \qquad f(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ 1 & \text{si } 0 < x \leq 1 \\ 0 & \text{si } 1 < x \end{cases}$$

1. INTRODUCTION

Considérons une série statistique (autrement dit une suite de nombres)

$$x_1 \quad x_2 \quad x_3 \ldots \quad x_k \ldots \quad x_n .$$

Vis-à-vis de cette série, on peut avoir deux attitudes :

a) Soit ces nombres sont considérés comme un échantillon extrait d'une population dont ils sont censés être plus ou moins représentatifs. Ils sont alors regardés comme les valeurs prises par n variables aléatoires X_1 , X_2 , ... X_k ... X_n de même loi et généralement indépendantes.

.../...

b) Soit la série est une entité que l'on étudie pour elle-même.

Exemple : a) On prend au hasard n arbres d'une plantation et l'on mesure leurs hauteurs.

b) On fait l'inventaire des salaires des agents du CIRAD.

Cette double attitude est si fréquente que, dans la plupart des logiciels, certains paramètres ne sont pas calculés de la même façon selon que l'on active un module de statistiques descriptives ou d'analyse de variance. Ainsi, pour la variance,

. dans un cas on calcule $\frac{1}{n} \sum_i (x_i - \bar{x})^2$

. dans l'autre $\frac{1}{n-1} \sum_i (x_i - \bar{x})^2$

Il n'est d'ailleurs pas impossible que l'on soit conduit à adopter vis-à-vis d'un échantillon donné les deux attitudes.

Une telle situation se retrouve dans l'étude des quantiles d'une série. On opposera alors "l'approche probabiliste" et "l'approche exploratoire" ou descriptive.

.../...

2. Approche probabiliste

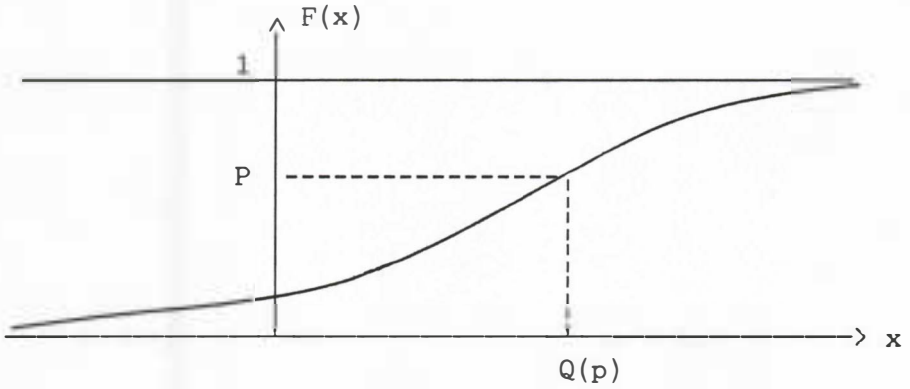
Si x_1 , x_2, ... x_k ... x_n est considéré comme un échantillon, cela signifie qu'il s'agit des valeurs prises par un n-uplet aléatoire

X_1 , X_2, ... X_k ... X_n

(les X_i sont supposés tous de même loi et indépendants).

Définition 1 : Soit X est une v.a. réelle dont on suppose la fonction de répartition $F(x) = p(X<x)$ continue et strictement croissante. Pour tout p ($0 \leq p \leq 1$) , on appelle quantile d'ordre p la racine Q(p) de l'équation : $F(x) = p$

$$Q(p) = F^{-1}(p)$$



En particulier, on note :

Quartile inférieur	Médiane	Quartile supérieur
$Q_1 = Q(1/4)$	$Q_2 = M = Q(1/2)$	$Q_3 = Q(3/4)$

On parle aussi de déciles Q(1/10) centiles Q(1/100) ...

Il est clair que ces paramètres induisent une ordonnance (plus grand que Q(p) ; plus petit que Q(p)) et qu'il est raisonnable d'avoir recours à l'échantillon ordonné

$X_{1:n}$ $X_{2:n}$ ... $X_{k:n}$ ... $X_{n:n}$

pour rechercher de bons estimateurs de Q(p) . On est donc conduit à s'interroger sur le comportement des $X_{k:n}$.

.../...

2.1. Loi des $X_{k:n}$

Il y a plusieurs façons d'aborder la loi de $X_{k:n}$

- soit l'étude directe de sa fonction de répartition $F_k(x)$
- soit l'étude d'une variable transformée.

a) Etude directe

$$F_k(x) = P(X_{k:n} < x) = \left\{ \begin{array}{l} \text{La probabilité que parmi les } n \text{ variables} \\ X_1 \quad X_2 \ \ldots \ X_k \ \ldots \ X_n \ , \ k \text{ d'entre elles} \\ \text{prennent des valeurs inférieures à } x \ . \end{array} \right\}$$

$$F_k(x) = \sum_{l=k}^{n} C_n^l \ (F(x))^l (1-F(x))^{n-l} = P(Z>k)$$

avec Z de loi binomiale de paramètres n et $F(x)$

Cette expression nécessite la connaissance de $F(x)$, elle est vraie quelle que soit la nature de X (continue ou discrète). La densité $f_k(x)$ s'écrit

$$f_k(x) = \frac{n!}{(n-k)!(k-1)!} F^{k-1}(x) \ (1-F(x))^{n-k} \ f(x)$$

Notons $B(p,q) = \int_0^1 t^{p-1} \ (1-t)^{q-1} \ dt$ et $I_u(p,q) = \dfrac{\int_0^u t^{p-1}(1-t)^{q-1}dt}{B(p,q)}$

On a $F_k(x) = I_{F(x)}(k,n-k+1)$

$I_u(p,q)$ a été tabulé (Karl PEARSON). Cela signifie que si l'on connaît $F(x)$ on peut calculer $P \{ X_{k:n} < x \}$ pour tous les couples (k,x) .

b) Etude indirecte

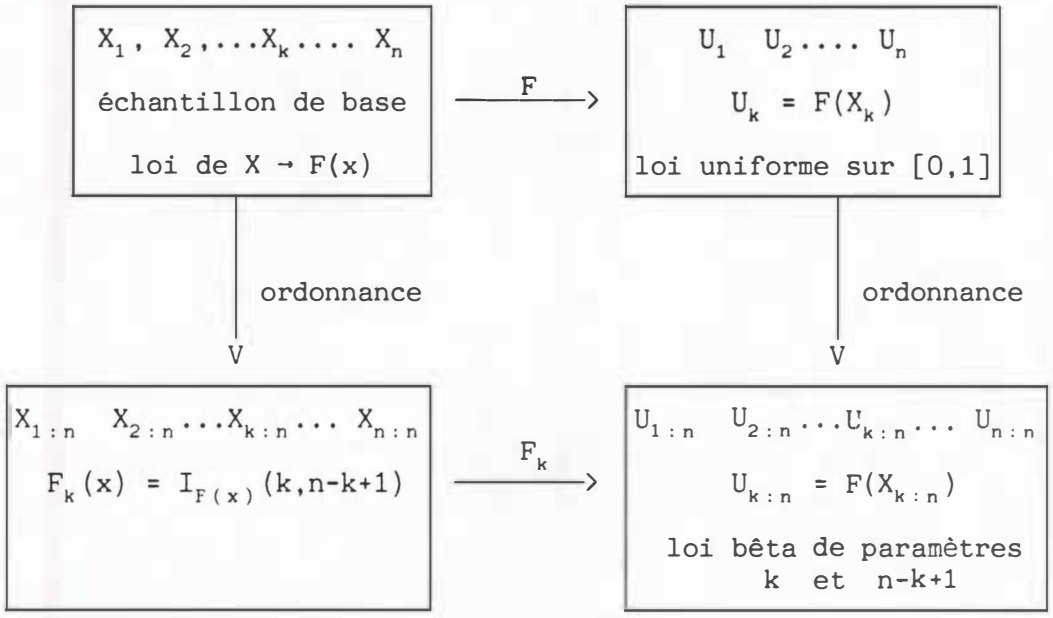
b1) F étant la fonction de répartition de X , la variable U égale à $F(X)$ ($U=F(X)$) suit une loi uniforme sur $[0,1]$. Ce résultat est classique :

$$P(U<u) = P \ (\ F(X)<u \ ) = P \ (\ X<F^{-1}(u) \ ) = F \ (\ F^{-1}(u) \ ) = u$$

Conclusion : $U_k = F(X_k)$ est de loi uniforme sur $[0,1]$

.../...

b2) Considérons $F(X_{k:n}) = U_{k:n}$. L'étude de cette transformation conduit à montrer que $U_{k:n}$ suit une loi bien connue des statisticiens ; il s'agit de la loi bêta de paramètres k et n-k+1 (voir en annexe).



Conclusion : On peut avoir l'illusion de contrôler la loi des $X_{k:n}$.

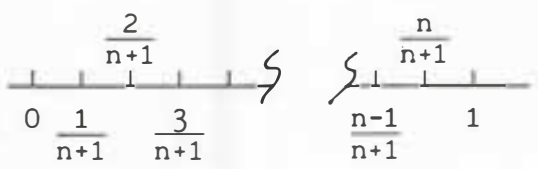
Remarque : U_k et $U_{k:n}$ prennent leurs valeurs dans [0,1].

c) **Les paramètres des variables $U_{k:n}$ et $X_{k:n}$**

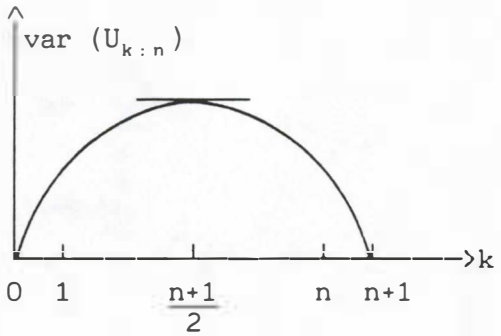
$U_{k:n}$ suit une loi bêta de paramètres (k,n-k+1) ; on en déduit

$$E(U_{k:n}) = \frac{k}{n+1} \qquad var(U_{k:n}) = \frac{(n-k+1)\,k}{(n+1)^2\,(n+2)}$$

Les espérances des $U_{k:n}$ divisent l'intervalle [0,1] en (n+1) intervalles de longueur k/n+1



Les $U_{k:n}$ n'ont pas tous la même variance



.../...

$$X_{k:n} = F^{-1}(U_{k:n})$$

on utilise un développement de F^{-1} au voisinage de $E(U_{k:n})$

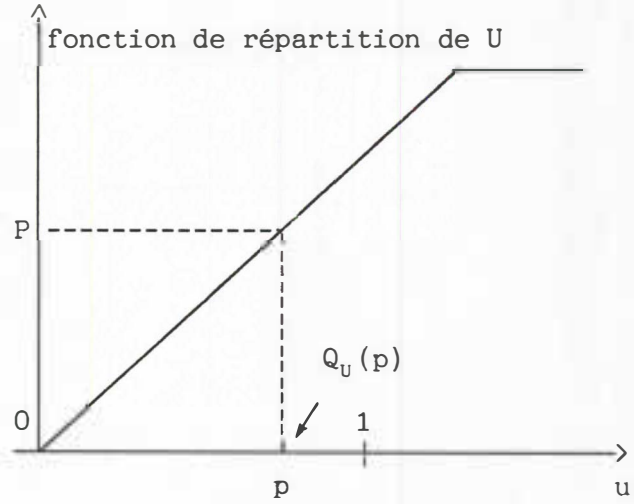
On peut aussi calculer les médianes des $U_{k:n}$ donc des $X_{k:n}$ (approché).

d) **L'intérêt à travailler sur les $U_{k:n}$**

Considérons une variable U qui suit une loi uniforme sur [0,1] ; on a (en utilisant la notation $Q_Y(p)$ pour le $p^{ème}$ quantile de Y) :

$$Q_U(p) = p \qquad [\ Q_X(p) = F^{-1}\ (\ Q_U(p)\ )\ ]$$

On sait situer $Q_U(p)$ par rapport aux espérances des $U_{k:n}$. On ne connaît par F mais on peut admettre que $X_{k:n}$ est une "bonne estimation" de F^{-1} ($E(U_{k:n})$). On peut donc positionner $Q_X(p)$ par rapport aux $X_{k:n}$.



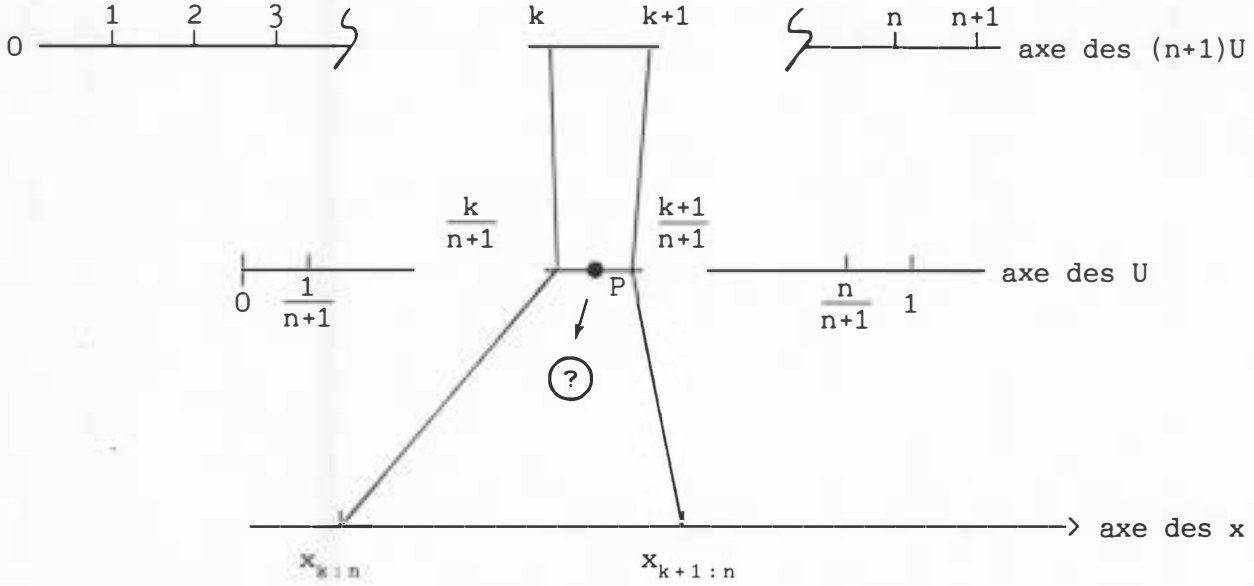
.../...

2.2. Estimation de Q(p)

Grossièrement, on peut dire que l'idée générale est de se positionner sur [0,1] ou sur [0 , n+1] (cette dernière façon évite d'avoir à traiter des valeurs fractionnaires k/n+1), puis à revenir par F^{-1} , ou plus exactement par la transformation

$$E(U_{k:n}) = \frac{k}{n+1} \longrightarrow x_{k:n} \quad \text{estimation de} \quad E(X_{k:n})$$

à Q(p)



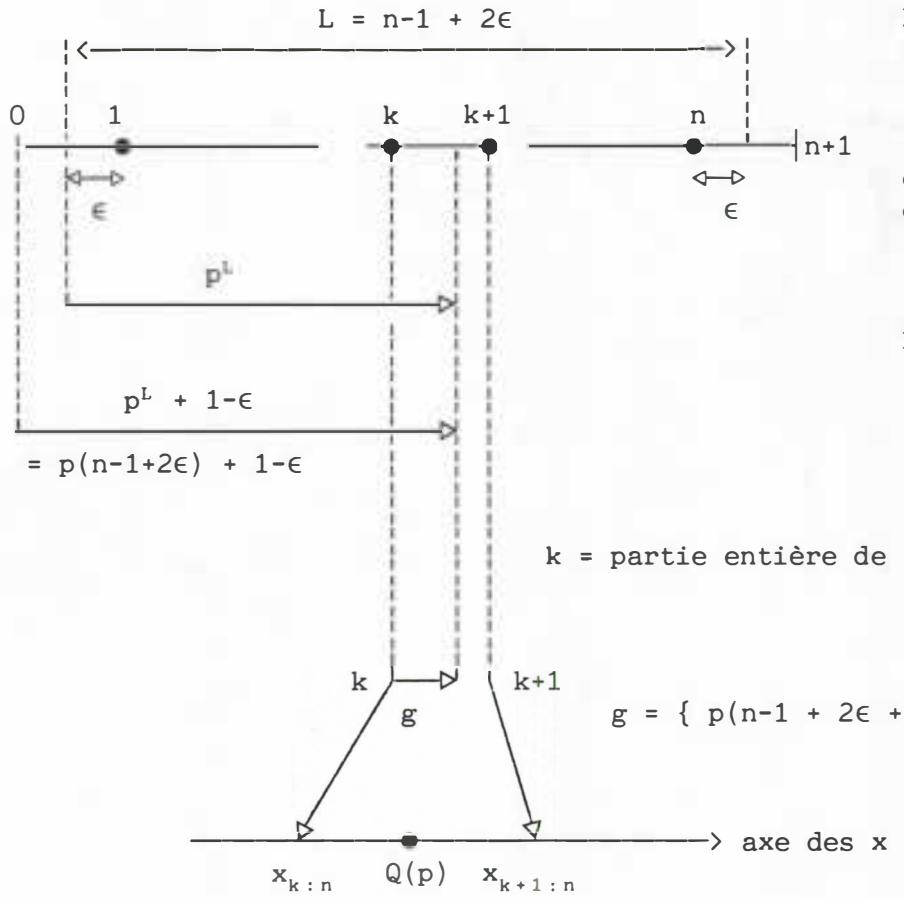
a) **Première approche**

On note k la partie entière de np et g le reste de la division de np par k

$$np = k + g \qquad g \in [0,1] \qquad k = [np]$$

Au lieu de considérer l'ensemble de l'intervalle [0,1] , de nombreux auteurs préconisent l'usage d'un intervalle tronqué à gauche et à droite. Il s'agit d'une pratique courante en recherche d'estimations robustes. Le schéma suivant a pour but d'expliciter la démarche suivie pour estimer Q(p) . On a choisi de travailler sur [0, n+1] pour plus de clarté.

.../...



Pour obtenir Q(p) , on effectue une interpolation $\Big\}$ $Q(p) = (1-g)\, x_{k:n} + g x_{k:n}$

Valeurs de $p(n + 2\epsilon - 1) + 1-\epsilon$ pour quelques valeurs de ϵ

ϵ	$p(n + 2\epsilon - 1) + 1-\epsilon$			
	p quelconque	p = 1/4	p = 1/2	p = 3/4
0	p(n-1) + 1	(n+3) /4	(n+1) /2	(3n+1) /4
1/3	p(n-1/3) + 2/3	(3n+7) /12	(n+1) /2	(6n+5) /12
1/2	pn + 1/2	(n+2) /4	(n+1) /2	(3n+2) /4
2/3	p(n+1/3) + 1/3	(3n+5) /12	(n+1) /2	(9n+7) /12
1	p(n+1)	(n+1) /4	(n+1) /2	(3n+3) /4

Remarque : La position de la médiane est toujours la même

si $n = 2m$, $M = 1/2(x_{m:n} + x_{m+1:n})$ si $n = 2m + 1$, $M = x_{m+1:n}$

Considérons la série

8 10 15 19 20 23 25 28 30 36 37 40

On obtient

	Q_1	M	Q_2
$\epsilon = 0$	18	24	31,5
$\epsilon = 1/2$	17	24	33
$\epsilon = 1$	16	24	30

$\epsilon = 1$ est utilisé dans la procédure Univariate de SAS.

Computational Methods

The sample mean, the sample standard deviation, the minimum, and the maximum are computed using the original data. All other statistics are computed after the data have been truncated to single precision (approximately seven significant digits).

Standard algorithms (Fisher 1973) are used to compute the moment statistics. Using the PCTLDEF= option, you can specify one of five methods for computing quantile statistics. See "SAS Descriptive Procedures" for computations.

Let n be the number of nonmissing values for a variable and let $x_1, x_2, \ldots, x_n$ represent the ordered values of the variable. For the tth percentile, where $p=t/100$, let

$$np = j + g$$

where j is the integer part and g is the fractional part of np.

The tth percentile, y, for example, is defined as:

DEFINITION 1: weighted average at x_{np}

$y = (1-g)x_j + gx_{j+1}$

where x_0 is taken to be x_1

DEFINITION 2: observation numbered closest to np

$y = x_i$

where i is the integer part of $np + 1/2$

DEFINITION 3: empirical distribution function

$y = x_j$ if $g = 0$
$y = x_{j+1}$ if $g > 0$

DEFINITION 4: weighted average aimed at $x_{p(n+1)}$

$y = (1-g)x_j + gx_{j+1}$

where $(n+1)p = j + g$

where x_{n+1} is taken to be x_n

DEFINITION 5: empirical distribution function with averaging

$y=(x_j+x_{j+1})/2$ if $g=0$
$y=x_{j+1}$ if $g>0$

where $np=j+g$.

.../...

2.3. Autres approches pour estimer Q(p)

a) **L'estimateur de Harrel et Davis**

$$Q(p) = \sum_{k=1}^{n} a_{k:n} X_{k:n} \qquad a_{k:n} = \frac{1}{B(np+1,n(1-p))} \int_{\frac{k-1}{n}}^{\frac{k}{n}} u^{np}(1-u)^{n(1-p)-1}\, du$$

b) **Jackknife**

On extrait un sous-échantillon de taille m de l'échantillon de base $(x_1 , x_2 , \ldots x_n)$.

On estime Q(p) par l'une des méthodes proposées.

On recommence un grand nombre de fois.

On obtient plusieurs estimations de Q(p)

$\hat{Q}_1(p)$ $\hat{Q}_2(p)$ .... $\hat{Q}_L(p)$

On calcule la moyenne et l'écart-type.

.../...

3. Approche exploratoire

Soit X une variable continue dont la fonction de répartition est strictement croissante. Notons Q_1 , Q_2 et Q_3 les quartiles de X .

A 25 % des valeurs prises par X sont inférieures à Q_1 .

B Chaque intervalle $]-\infty,Q_1]$ $[Q_1,Q_2]$ $[Q_2,Q_3]$ et $[Q_3,+\infty[$ contient 25 % de la population.

C $[Q_1,Q_3]$ contient 50 % de la population.

Ces différentes propriétés sont attachées au concept de quartiles et ceci souvent à tort lorsqu'on considère une distribution quelconque ou une série statistique.

Exemple : 1 2 $[Q_1=3]$ 4 5 6 7 8 9 et 2/9 = 0,22

Cela a conduit certains auteurs à rechercher des paramètres différents des quartiles mais plus aptes à être interprétés selon les critères A B et/ou C .

Définition : Considérons une série ordonnée

$$x_{1:n} \quad x_{2:n} \quad \cdots \quad x_{k:n} \quad \cdots\cdots \quad x_{n:n}$$

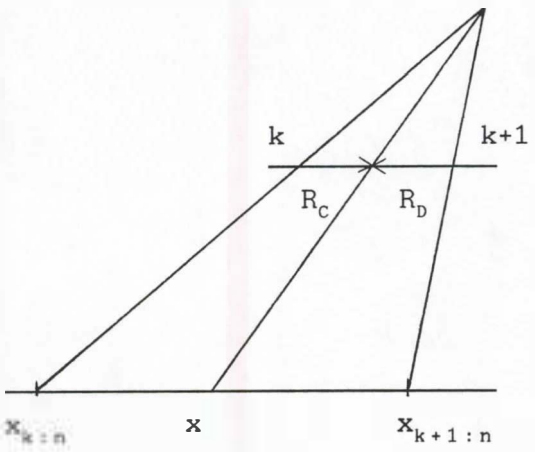
$x_{k:n}$ a un rang $R_C = k$ dans le sens croissant et $R_D = n-k+1$ dans le sens décroissant. On a

$$R_C + R_D = n+1$$

On appelle profondeur (*depth*) de $x_{k:n}$ la plus petite des valeurs R_C , R_D

$$pr(x_{k:n}) = \inf(R_C, R_D)$$

Cette notion de profondeur peut être généralisée à un nombre quelconque n'appartenant pas à la série



On opère par homothétie

.../...

Ainsi la médiane a pour profondeur $\frac{n+1}{2}$

En utilisant cette notion de profondeur, on définit :

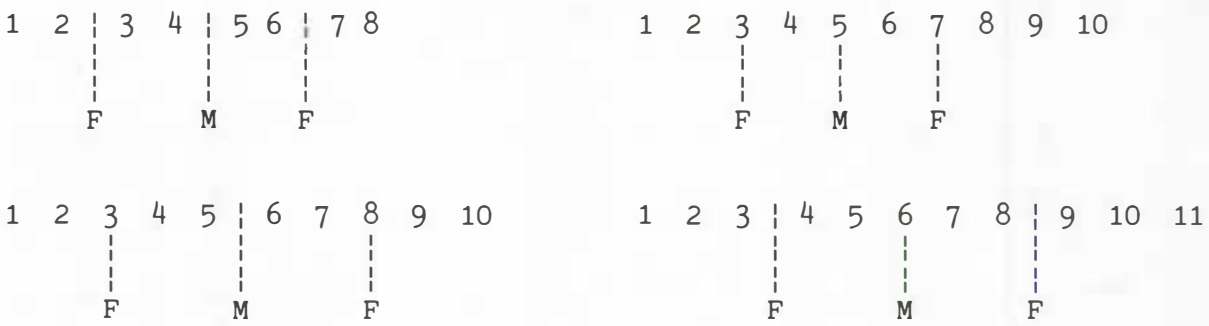
les quatrièmes F $pr(\text{quatrième}) = \frac{pr(\text{médiane})+1}{2}$

les huitièmes E $pr(\text{huitième}) = \frac{pr(\text{quatrième})+1}{2}$

et ainsi de suite. Ces paramètres sont notés par des lettres

M	F	E	D	C	B	A	Z	Y	X
1/2	1/4	1/8	1/16	1/32	1/64	1/128	1/256	1/512	1/1024

Exemple :



.../...

DISTRIBUTION BETA

Une variable Y suit une loi bêta de paramètres p et q si sa densité est de la forme

$$f(y) = \frac{1}{B(p,q)} \frac{(y-a)^{p-1}(b-y)^{q-1}}{(b-a)^{p+q-1}} \quad \text{si } a \leq y \leq b \; ; \; f(y) = 0 \quad \text{ailleurs}$$

p et q sont deux nombres positifs.

$$B(p,q) = \frac{\Gamma(p/2)\,\Gamma(q/2)}{\Gamma((p+q)/2)} \qquad \text{et} \quad \Gamma(p) = \int_0^{+\infty} e^{-x}\, x^{p-1}\, dx$$

Remarque : La fonction gamma Γ est telle que

$$\Gamma(p) = (p-1)\Gamma(p-1) \qquad \Gamma(1/2) = \sqrt{\pi}$$

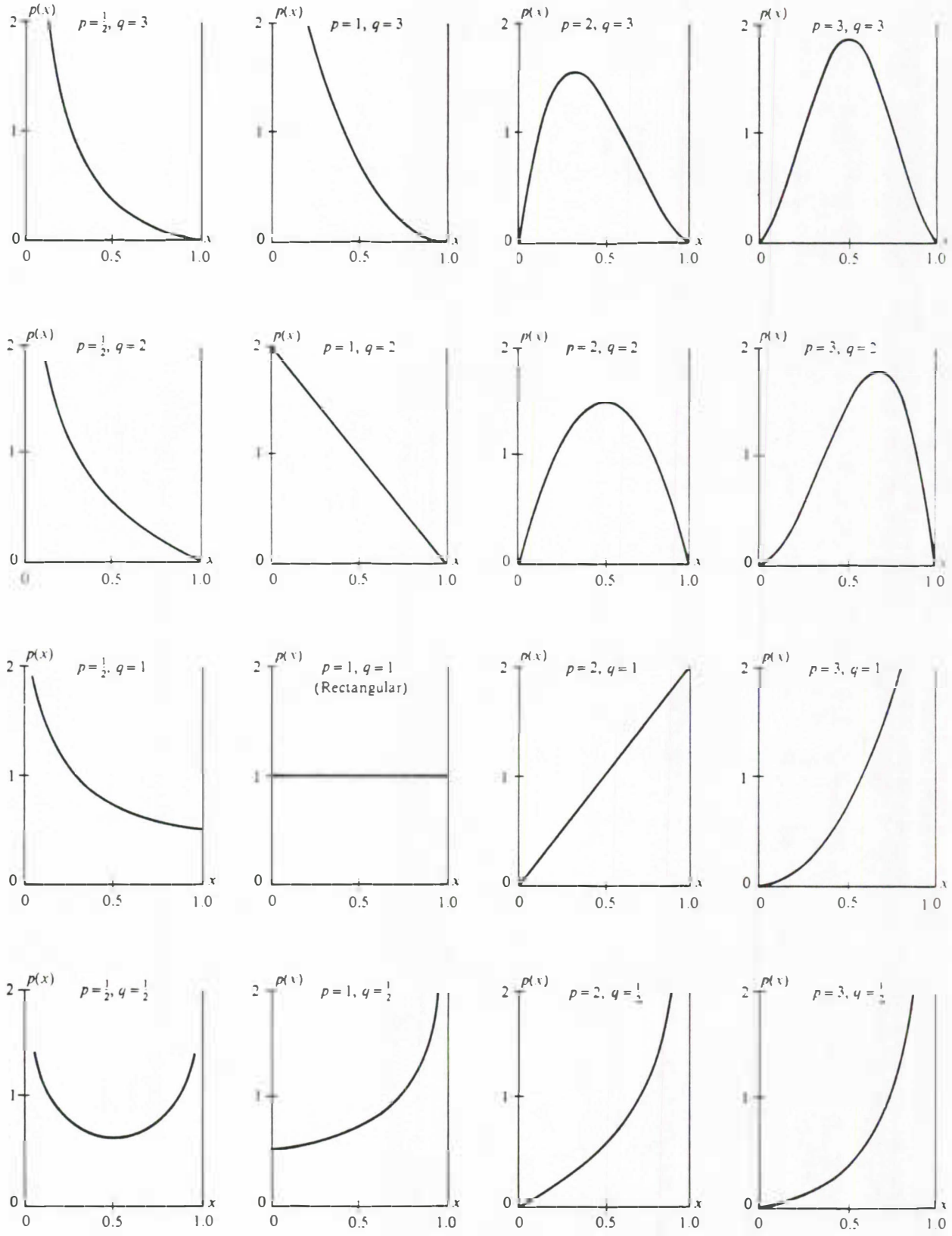
si p est entier $\Gamma(p) = (p-1)\,!$

Si l'on fait le changement de variable $X = \frac{Y-a}{b-a}$, on obtient une forme dite standard, avec

$$f(x) = \frac{1}{B(p,q)}\, x^{p-1}(1-x)^{q-1} \quad \text{si} \quad 0 \leq x \leq 1 \qquad E(X) = \frac{p}{p+q} \qquad \text{var}X = \frac{pq}{(p+q)^2(p+q+1)}$$

Cette distribution est intéressante dans la mesure où sa densité peut prendre un grand nombre de formes comme on peut l'apprécier sur le graphique ci-après.

.../...



Beta Density Functions

LES QUANTILE-QUANTILE PLOTS (QQ PLOTS)

Septembre 1992

Matthieu LESNOFF

BIOMETRIE
CIRAD - Forêt

LES QUANTILE-QUANTILE PLOTS (QQ PLOTS)

Le QQ plot est une méthode graphique très efficace pour comparer les distributions de deux ensembles de données. Elle correspond à la représentation des quantiles d'une première distribution en fonction des quantiles du même ordre d'une deuxième distribution :

> Soit X et Y deux variables aléatoires, Fx et Fy leur fonction de répartition et Qx(p) et Qy(p) leurs quantiles d'ordre p ($Fx(Qx(p))=P(X<Qx(p))=p$ et $Fy(Qy(p))=P(Y<Qy(p))=p$). Le Q-Q plot (X,Y) est alors la représentation de Qy(p) en fonction de Qx(p).

Deux formes de QQ plots se présentent : le QQ plot empirique (Wilk et Gnanadesikan, 1968) qui compare deux distributions empiriques ("les deux échantillons observés sont-ils générés par une même loi ?), et le QQ plot théorique, qui remplace l'une des distributions par une distribution théorique fixée ("l'échantillon observé est-il généré par une loi de probabilité connue ?).

Si les deux distributions sont identiques, les points du graphique suivent exactement la droite y=x (bissectrice). Les écarts à cette droite donnent une information détaillée sur la différence entre ces distributions :

- Comparaison des localisations (paramètres centraux) et des dispersions des distributions.

- Comparaison de la forme des distributions. Si les deux distributions sont de même famille (deux lois normales $N(\mu_1,\sigma_1)$ et $N(\mu_2,\sigma_2)$ par exemple), le QQ plot est linéaire.

Dans certains cas (favorables), le QQ plot permet de quantifier graphiquement l'écart entre les deux distributions.

I. QQ PLOTS EMPIRIQUES.

1. Définition des quantiles empiriques.

Chambers et al (1983) et Cleveland (1985) définissent le quantile empirique Q(p) d'ordre p de la manière suivante :

> Soit un échantillon {xi} i=1,n que l'on trie par ordre croissant en {x(i)} i=1,n ($x(1)\leq x(2)\ldots\leq x(n)$). On définit la probabilité ph=(h-0.5)/n. On cherche alors Qx(ph) tel que h=n.ph+0.5.
>
> On pose h=k+f avec k=[h] ([ ] : partie entière)
> alors Q(ph)=(1-f).Q(pk) + f.Q(pk+1) avec Q(pk)=x(k) et Q(pk+1)=x(k+1)
> (par convention Q(p)=x(1) si p<p1 et Q(p)=x(n) si p>pn).

D'après cette définition, Qx(pi)=x(i) pour toutes les probabilités pi=(i-0.5)/n i=1,n. Les quantiles Qx(ph) avec h≠i sont calculés par interpolation linéaire (fig.1).

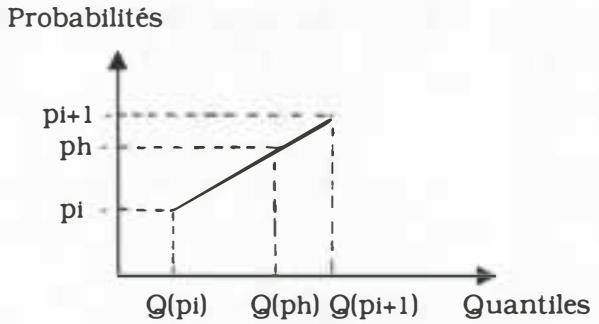


Fig.1 : Calcul des quantiles empiriques par interpolation linéaire.

Exemple : calcul de la médiane Qx(0.5) . n pair : Qx(0.5)=(x(n/2)+x((n/2)+1))/2
. n impair : Qx(0.5)=x((n+1)/2)

2. Construction du QQ plot empirique.

Soit deux échantillons comparés {xi} i=1,n et {yj} j=1,m triés par ordre croissant en {x(i)} i=1,n et {y(j)} j=1,m.

- si les deux échantillons sont de même taille (n=m) : pour tout i Qx(pi)=x(i) et Qy(pi)=y(i). Le QQ plot est simplement la représentation de y(i) en fonction de x(i) i=1,n.

- si les deux échantillons sont de taille inégale (m≠n) : les quantiles d'un des deux échantillons devront être calculés par interpolation. Par convention (Chambers et al, 1983), les valeurs du plus petit des deux échantillons sont conservées entières : soit m<n alors les quantiles Qy(pj)=y(j) (avec pj=(j-0.5)/m) sont représentés en fonction des quantiles interpolés Qx(pj)=x*(j).
Soit h tel que (h-0.5)/n=(j-0.5)/m, donc h=(n/m)(j-0.5)+0.5. On pose h=k+f avec k=[h], alors Qx(pj)=x*(j)=(1-f).x(k)+f.x(k+1).

3. Interprétation du QQ plot.

3.1 Exemples.

3.1.1 Exemple n°1.

Chambers et al (1983) comparent les distributions de températures relevées dans deux villes américaines (Newark et Lincoln) de 1963 à 1974 (fig.2). Les deux échantillons de températures sont de même taille et le QQ plot correspond au graphique y(i)=f(x(i)). Les distributions de températures apparaissent très similaires pour les températures supérieures à 50°F, puisque les points suivent la bissectrice. En dessous de 50°F, les températures deviennent plus fortes pour la ville de Newark.

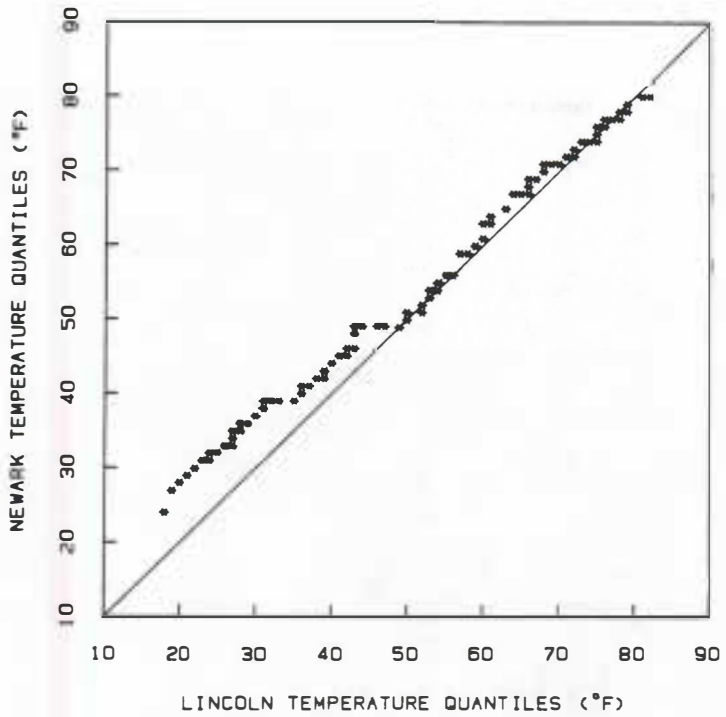


Fig.2 : QQ plot des températures de Newark et Lincoln entre 1963 et 1974. La droite représentée en trait plein correspond à la bissectrice (D'après Chambers et al., 1983).

Sur la période d'étude, les périodes chaudes (étés) ont donc tendance à correspondre à des températures semblables pour les deux villes, alors que les périodes froides atteignent des températures généralement plus faibles (d'environ 8°F) pour Lincoln que pour Newark.

Il faut bien comprendre la différence entre un graphe de dispersion (x,y) et un QQ plot. Le graphe de dispersion est utilisé pour rechercher une relation systématique entre x et y : "les températures de Lincoln sont-elles liées à celles de Newark ?". Le QQ plot est utilisé pour comparer deux distributions : "pendant la période d'étude, les temps chauds de Lincoln correspondent-ils généralement à des températures plus fortes ou plus faibles que celles des temps chauds de Newark ?"

3.1.2 Exemple n°2.

Le QQ plot de la figure n°3 permet de comparer les distributions de concentrations en ozone dans les villes de Stamford et de Yonkers. Les deux échantillons étant de taille inégale, les quantiles de Yonkers sont calculés par interpolation. La concentration en ozone de Stamford apparaît toujours plus forte que celle de Yonkers, et cette différence augmente d'autant plus que les valeurs deviennent élevées : plus les concentrations deviennent fortes, plus les écarts absolus deviennent grands. L'aspect linéaire du QQ plot (les points suivent approximativement une droite de pente 1.6) permet de quantifier la différence entre les deux distributions : chaque quantile de Stamford est environ 1.6 fois plus fort que le quantile correspondant de Yonkers. Globalement et pendant la période d'étude, la différence de concentration en ozone entre Stamford et Yonkers a donc été de 60% (mais ceci ne signifie pas que pour un jour donné la concentration en ozone soit nécessairement plus forte à Stamford qu'à Yonkers).

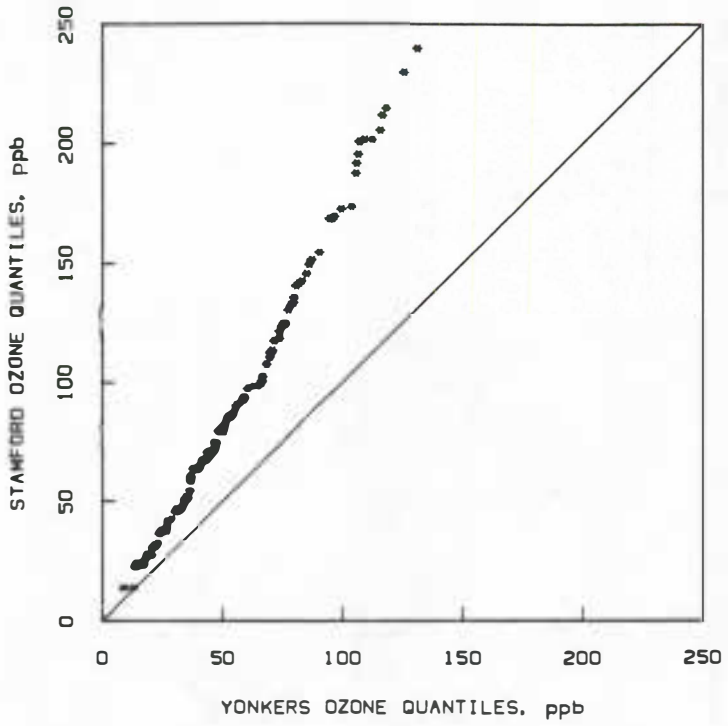


Fig.3 : QQ plot des concentrations en ozone de Stamford et de Yonkers. La droite représentée en trait plein correspond à la bissectrice (D'après Chambers et al., 1983).

3.2 Formes classiques et interprétations.

3.2.1 QQ plots linéaires.

Les QQ plots linéaires sont de la forme : y(i)≈k.x(i)+c (fig.4 à fig.7). Cette relation permet d'avoir une estimation informelle des paramètres de localisation (médiane, moyenne, etc.) et de dispersion (intervalle interquartile, écart-type, etc.) de la loi Py en fonction des paramètres de Px :

. localisation(Y) ≈k.localisation(X)+c
. dispersion(Y)≈k.dispersion(X)

Fig.4 : y(i)=x(i)+c. Les distributions diffèrent uniquement par leur paramètre de localisation ("globalement, X est plus faible que Y").

Fig. 5 : y(i)=k.x(i). Les distributions diffèrent par leur localisation et leur dispersion ("globalement X est plus faible que Y, et l'écart augmente d'autant plus que X et Y sont grands").

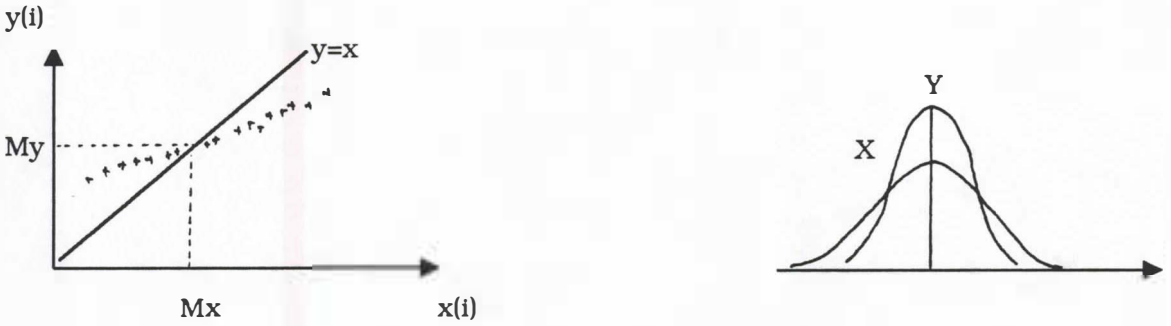


Fig.6 : y(i)=k.x(i)+c avec médiane(X)=médiane(Y). X et Y ont même localisation, mais Y est moins dispersée. Les faibles valeurs de Y sont plus grandes que les faibles valeurs de X, et inversement pour les grandes valeurs.

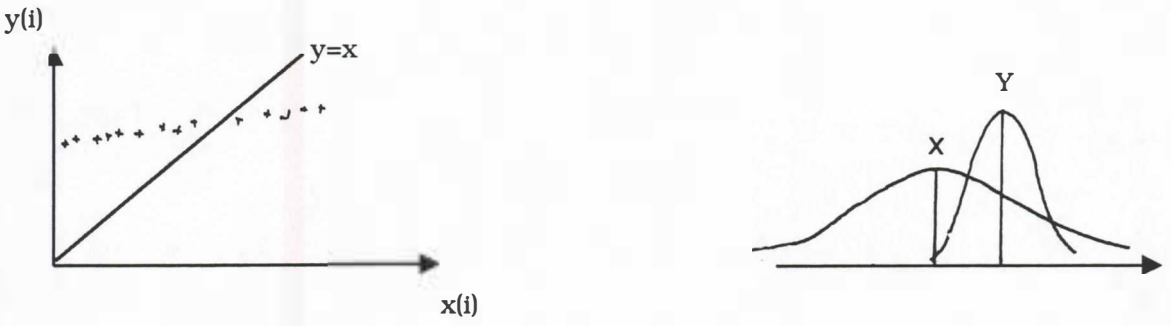


Fig.7 : y(i)=k.x(i)+c avec médiane(X)≠médiane(Y). L'interprétation est la même que celle de la figure n°6, mais les distributions n'ont pas même localisation.

3.2.2 QQ plots non linéaires.

Dans le cas non linéaire, en plus de différer par leur localisation ou leur dispersion, les deux distributions n'ont plus la même forme ; elles n'appartiennent plus à une même famille de lois. Les figures n°8 à n°11 présentent quatre schémas classiques de non linéarité.

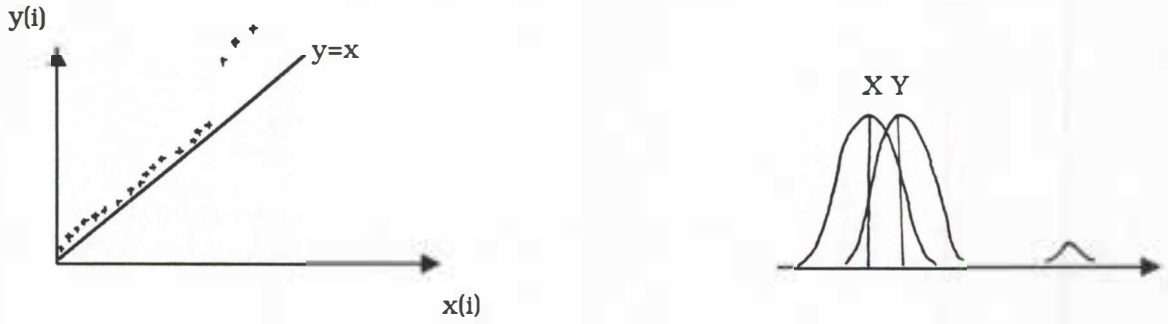


Fig.8 : Présence de valeurs extrêmes ("outliers") dans la distribution de Y.

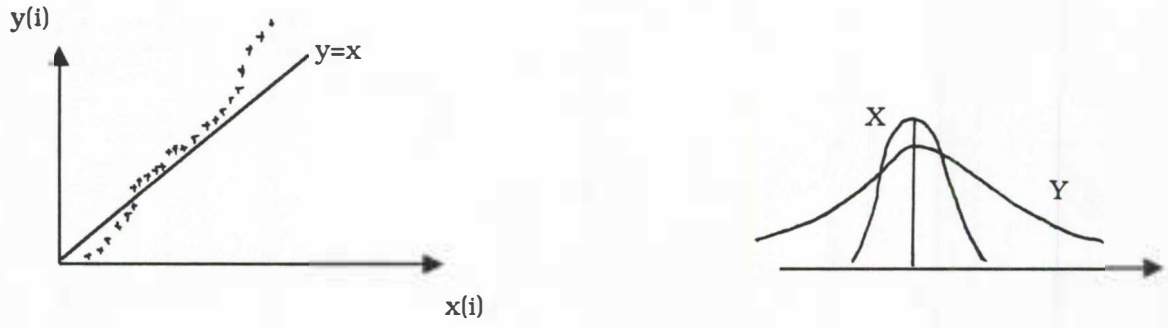


Fig.9 : La distribution de Y a des queues plus lourdes que la distribution de X.

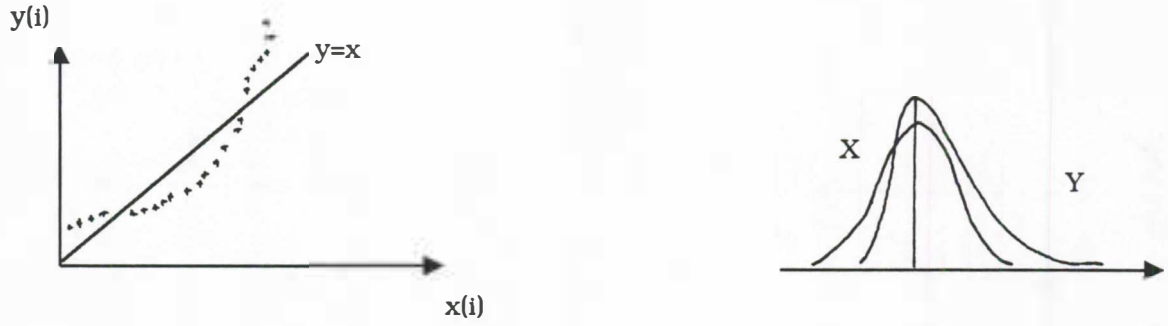


Fig.10 : La distribution de Y est dissymétrique (étalement à droite).

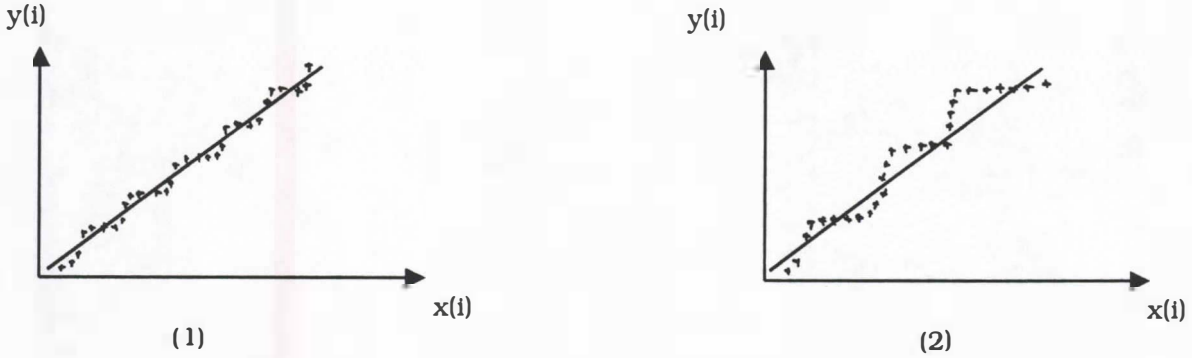


Fig.11 : Présence de plateaux sur le QQ plot. Ces plateaux peuvent provenir de problèmes d'arrondis (1) ou de concentrations de points dues à la planification de l'échantillonnage (2).

3.2.3 Remarques sur la construction des QQ plots.

- En cas de grands échantillons, le problème de chevauchement des points sur le graphique intervient et nuit à la lisibilité du QQ plot. Une solution est de réduire a posteriori la taille des échantillons comparés en éliminant systématiquement un certain nombre de données dans les échantillons ordonnés (par exemple en ne conservant qu'une donnée sur trois).

- En plus d'une procédure qui trace directement la bissectrice, il est très utile de disposer d'une procédure de lissage et/ou de régression des points du QQ plot pour en apprécier la forme (et les différents paramètres dans le cas linéaire).

3.3 Compléments des QQ plots.

Il est possible de construire quelques graphiques complémentaires pour faciliter l'analyse des écarts entre les deux distributions. Les deux premiers graphiques présentés (fig.12 et fig.13) correspondent aux représentations de y(i)-x(i) et de y(i)/x(i) en fonction de x(i).

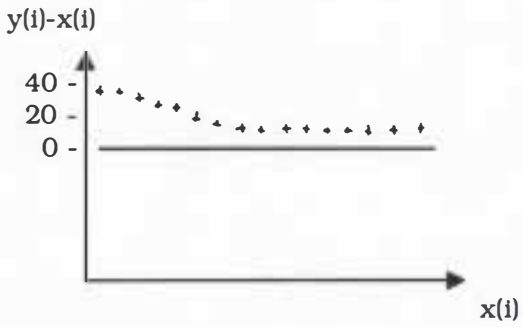


Fig.12 : Graphique y(i)-x(i)=f(x(i)). L'écart entre les quantiles est d'environ 40 unités pour les faibles valeurs de X et d'environ 10 unités pour les valeurs les plus fortes.

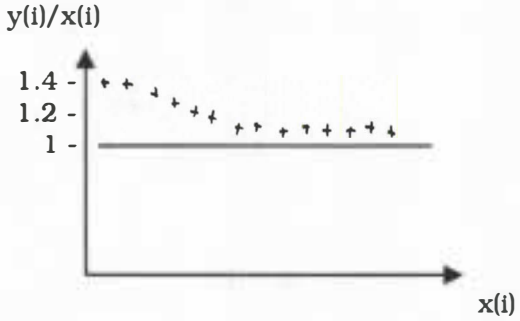


Fig.13 : Graphique y(i)/x(i)=f(x(i)) (il ne peut être construit que si les quantiles sont tous positifs). L'écart entre les quantiles est d'environ 40% pour les faibles valeurs de X et d'environ 10% pour les plus grandes valeurs.

Ces deux graphiques ne traitent pas X et Y de manière symétrique. J.W. Tukey a développé un graphique qui élimine ce problème en représentant y(i)-x(i) en fonction y(i)+x(i) (fig.14) : le "Tukey sum-difference graph" (Cleveland, 1985). Il s'interprète comme les deux graphiques précédents.

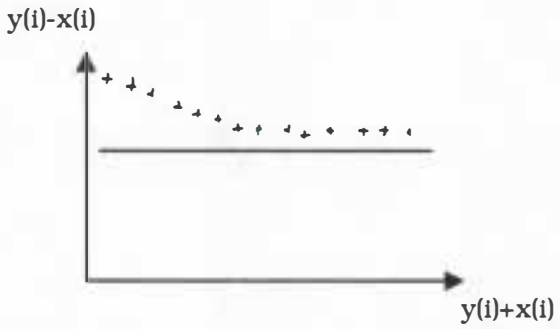


Fig.14 : Graphique y(i)-x(i)=f(y(i)+x(i)) ("Tukey sum-difference graph"). Les quantiles de Y sont supérieurs à ceux de X pour les faibles valeurs de X et de Y, et inversement pour leurs grandes valeurs.

II. QQ PLOTS THEORIQUES.

1. Construction du QQ plot.

L'objet du QQ plot théorique est de comparer une distribution empirique Y (observée) à une distribution théorique prédéfinie. La première phase de la construction du graphique est le tri de l'échantillon étudié {yi} i=1,n en {y(i)} i=1,n , chaque y(i) représentant le quantile empirique Qe(pi) d'ordre pi pour toutes les probabilités pi=(i-0.5)/n.
La deuxième phase est le calcul des quantiles théoriques Qt(pi) correspondant à la loi de probabilité choisie pour chaque probabilité pi : $Qt(pi)=F^{-1}(pi)$, F étant la fonction de répartition de la loi théorique. Le QQ plot théorique est alors la représentation de Qe(pi) en fonction de Qt(pi). Chambers et al (1983) donnent les formules de calcul des quantiles de plusieurs lois courantes. Certaines de ces lois ont des fonctions de répartition facilement inversibles menant à des formules de calcul exactes des quantiles (tab.1). Pour les autres lois, seules des approximations sont

Information for quantile-quantile plot construction — distributions in closed form. p_i is equal to $(i-.5)/n$.

	Plot Construction		What is Estimated by	
Family	*Ordinate*	*Abscissa*	*Intercept*	*Slope*
Uniform	$y_{(i)}$	p_i	μ	λ
One Parameter Exponential	$y_{(i)}^{1/3}$	$(-\log_e(1-p_i))^{1/3}$	0	λ
Two Parameter Exponential	$y_{(i)}$	$-\log_e(1-p_i)$	μ	λ
Weibull	$\log_e(y_{(i)})$	$\log_e(-\log_e(1-p_i))$	$\log_e\lambda$	θ^{-1}

Tab.1 : Construction du QQ plot théorique dans le cas de lois à fonction de répartition inversible (abscisses : quantiles théoriques Qt(pi)=F^{-1}(pi) ; ordonnées : quantiles empiriques Qe(pi)=y(i)). Les paramètres de localisation (μ) et de dispersion (σ) de la distribution empirique Y sont estimés graphiquement par l'ordonnée du point d'abscisse nulle et par la pente de la droite Qe(pi)=f(Qt(pi)) (D'après Chambers et al., 1983).

	Plot Construction		What is Estimated by	
Family	*Ordinate*	*Abscissa*	*Intercept*	*Slope*
Normal	$y_{(i)}$	$\Phi^{-1}(p_i)$	μ	σ
Power Normal (Including Log Normal)	$y_{(i)}^{(\theta)}$	$\Phi^{-1}(p_i)$	μ	*
Gamma	$y_{(i)}^{1/3}$	$[G_\alpha^{-1}(p_i)]^{1/3}$	0	$\lambda^{1/3}$
Chi-square	$y_{(i)}^{1/3}$	$[2\,G_{\alpha/2}^{-1}(p_i)]^{1/3}$	0	$\lambda^{1/3}$
Half-Normal	$y_{(i)}$	$\Phi^{-1}(p_{(i)}/2+.5)$	0	σ

p_i is equal to $(i-.5)/n$.

Normal with mean 0 and variance 1

$$\Phi^{-1}(p) \approx [c-(2.30753+.27061c)/(1+.99229c+.04481c^2)]\text{sign}(p-.5)$$

where

$$c = (-2\log(\min(p,1-p)))^{1/2}$$

$$\text{sign}(x) = \begin{cases} +1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0. \end{cases}$$

Cube-root gamma with shape parameter α, and λ = 1

$$[G_\alpha^{-1}(p)]^{1/3} \approx \alpha^{1/3}\left[1 - \frac{1}{9\alpha} + \frac{c}{3\sqrt{\alpha}}\right]$$

where,

$$c = \Phi^{-1}(p).$$

Tab.2 : Construction du QQ plot théorique dans le cas de lois à fonction de répartition non inversible (abscisses : quantiles théoriques Qt(pi)=F^{-1}(pi) ; ordonnées : quantiles empiriques Qe(pi)=y(i)). Les paramètres de localisation (μ) et de dispersion (σ) de la distribution empirique Y sont estimés graphiquement par l'ordonnée du point d'abscisse nulle et par la pente de la droite Qe(pi)=f (Qt(pi)) (D'après Chambers et al., 1983).

fournies (tab.2). Dans le cas de certaines lois dissymétriques, Chambers et al conseillent d'utiliser une échelle "racine cubique" : Qe(pi) $^{1/3}$=f(Qt(pi) $^{1/3}$). Cette transformation symétrise les données et améliore la lisibilité du graphique en évitant l'accumulation de points près de l'origine.

2. Interprétation du QQ plot théorique.

Les règles d'interprétation du QQ plot théorique sont les mêmes que celles du QQ plot empirique :

> Si le graphique a un aspect relativement linéaire, le modèle théorique proposé s'accorde bien avec la distribution empirique. La loi de Y (empirique) peut être exprimée en fonction de la loi de X (théorique) sous la forme : Y≈kX+c avec localisation(Y)≈k.localisation(X)+c et dispersion(Y)≈k.dipersion(X).

Le QQ plot théorique ne permet donc pas simplement de comparer une distribution empirique à une seule distribution théorique, mais plutôt à un ensemble de distributions de même famille. Par exemple, le QQ plot théorique construit en utilisant les quantiles théoriques d'une loi normale N(0,1) est suffisant pour tester l'adéquation de n'importe quelle loi N(μ,σ) :

> Si X suit une loi normale N(0,1), alors Y=σX+μ suit une loi normale N(μ,σ). L'ordonnée du point d'abscisse nulle et la pente de la droite du QQ plot sont des estimations informelles de la moyenne et de l'écart-type de Y (fig.15) ; c'est le principe de la "droite de Henry". Les tableaux n°1 et n°2, ainsi que Callot (1984, p149-173) et Lecoutre et Tassi (1987, p 321-326) donnent la signification des paramètres de la droite pour les principales lois classiques.

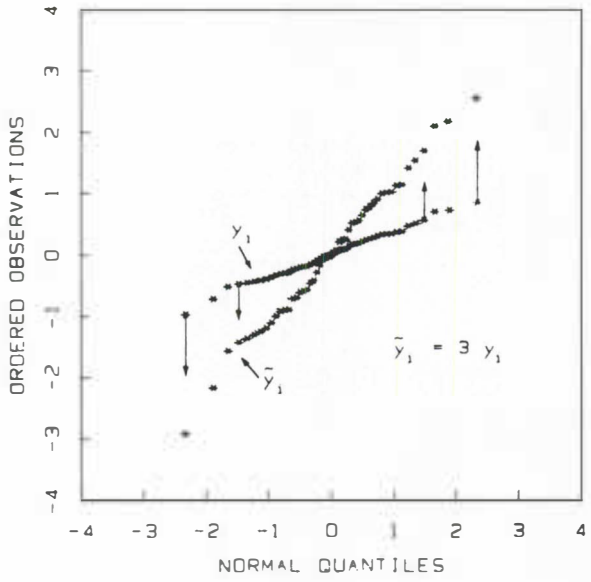


Fig.15 : QQ plot Normal pour les distributions empiriques Y et $\tilde{Y}$ (Y≈(1/3)$\tilde{Y}$). $\tilde{Y}$ suit une loi N(0,1) alors que Y suit une loi N(0,1/3) (D'après Chambers et al., 1983).

3. Cas de distributions ayant des paramètres de forme inconnus.

L'avantage du QQ plot mettant en jeu la loi normale est qu'il est construit sans avoir besoin d'estimer les paramètres caractéristiques de cette loi (la loi normale est totalement définie par sa moyenne et sa variance). Certaines familles de lois ont cependant d'autres paramètres qu'il faut estimer avant de pouvoir construire le QQ plot.

3.1 La distribution Gamma.

La loi Gamma(α) de densité $f(y;\alpha,\beta)=\beta^{-\alpha}y^{\alpha-1}\exp(-y/\beta)/G(\alpha)$ (β paramètre de dispersion et G fonction gamma) a un paramètre de forme α qui doit être spécifié pour le calcul des quantiles théoriques. Le problème est que la procédure d'estimation de α qui précède la construction du QQ plot est basée sur l'hypothèse que la distribution étudiée appartient à la famille de lois choisie, or c'est justement ce que l'on cherche à savoir : le paramètre α peut très bien être estimé de manière précise si on accepte le modèle Gamma, alors qu'en fait les données ne suivent pas ce modèle.Par ailleurs, la robustesse du QQ plot comme outil d'exploration de ce modèle de loi dépend directement de la robustesse de la procédure d'estimation du paramètre α : si les données sont contaminées par quelques valeurs extrêmes, celles-ci peuvent avoir une grande influence sur l'estimation et donc sur la forme finale du QQ plot (le graphique peut avoir un aspect non linéaire alors que le modèle Gamma était adapté)

D'après Chambers et al (1983), une bonne stratégie est d'une part d'utiliser des estimateurs robustes de α, et d'autre part de construire plusieurs QQ plots pour un nombre raisonnable de valeurs de α (et non se limiter à une seule valeur), par exemple en balayant l'intervalle de confiance de l'estimation. L'examen des différents QQ plots construits d'après cette fourchette de valeurs permettra de juger si le modèle Gamma peut s'appliquer raisonnablement à la distribution observée : l'un des QQ plots est-il linéaire ?

3.2 QQ plots et transformations puissances.

Dans de nombreux cas de distributions dissymétriques, l'application d'une transformation puissance du type :

$$x^{(p)} = \begin{cases} . \; x^p \text{ si } p > 0 \\ . \; \log(x) \text{ si } p = 0 \\ . \; -x^p \text{ si } p < 0 \end{cases}$$

permet d'améliorer la symétrie de l'ensemble. Le problème est de déterminer la valeur $p=p^*$ qui symétrise au mieux les données. p^* peut être estimée de manière numérique (Box and Cox, 1964), mais aussi de manière visuelle grâce au QQ plot (fig.16) :

> Si p^* est la valeur qui symétrise les données, alors une transformation puissance $x^{(p)}$ avec $p>p^*$ entraîne une dissymétrie à droite, et inversement pour une transformation $x^{(p)}$ avec $p<p^*$. Il est ainsi possible d'avoir une idée de p^* en diminuant de manière progressive l'intervalle $[p_1,p_2]$ (avec $p_1<p^*$ et $p_2>p^*$) jusquà l'obtention d'un QQ plot linéaire.

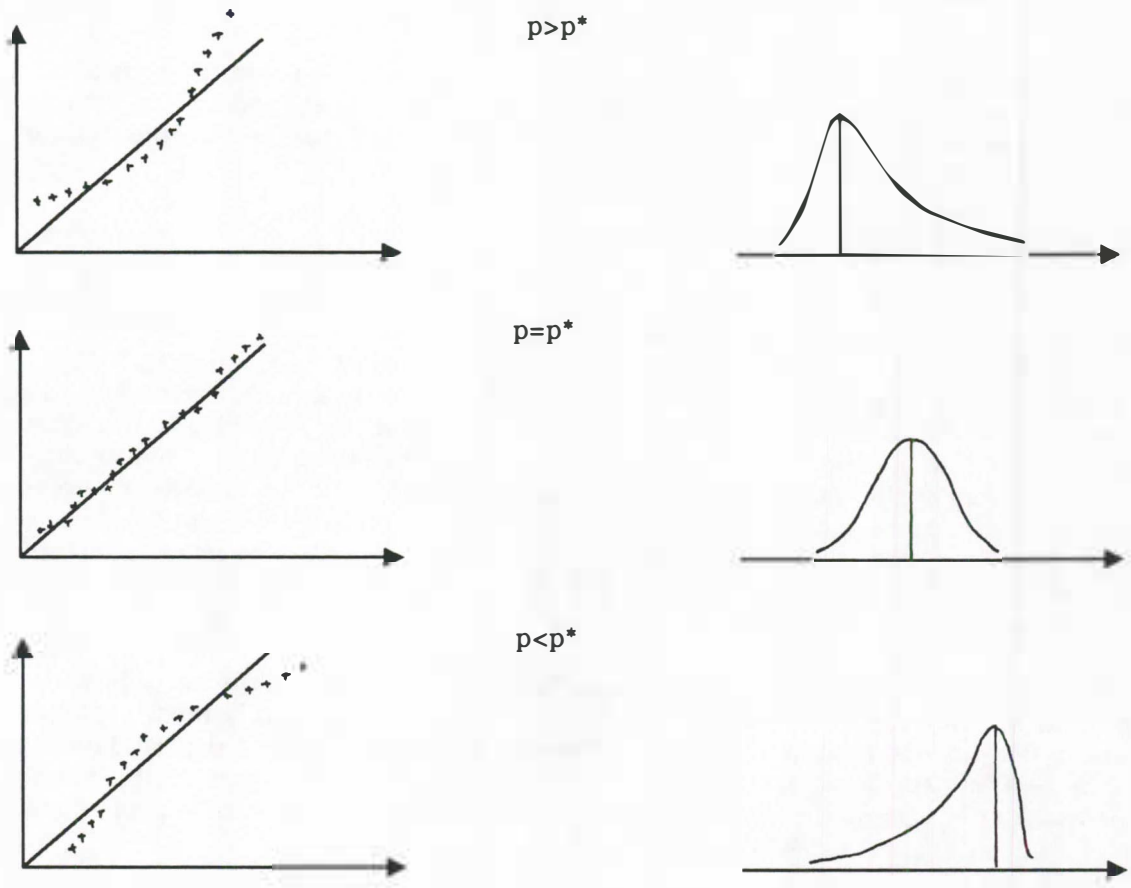


Fig.16 : Effets de transformations puissances x^p sur la forme d'une distribution.

BIBLIOGRAPHIE.

- BOX (G.E.P.), COX (D.R.), 1964. An analysis of transformation. JRSS, B, 26:211-252.

- CALLOT (G.), 1984. Cours de statistique descriptive. Dunod, 488 p.

- CHAMBERS (J.M), CLEVELAND (W.S.), KLEINER (B.), TUKEY (P.A.), 1983. Graphical methods for data analysis. Wadsworth International Group, 395 p.

- CLEVELAND (W.S.), 1985. The element of graphical data. Wadsworth Advanced Books and Software, 323 p.

- LECOUTRE (J.P.), TASSI (P.), 1987. Statistique non paramètrique et robustesse. Economica, 455 p.

- WILK (M.B.), GNANADESIKAN (R.), 1968. Probability plotting methods for the analysis of data. Biometrika, 55:1-17.

TECHNIQUES GRAPHIQUES ET TRANSFORMATIONS PUISSANCES

Septembre 1992

Matthieu LESNOFF

BIOMETRIE
CIRAD - Forêt

TECHNIQUES GRAPHIQUES ET TRANSFORMATIONS PUISSANCES.

Il est en général possible d'éliminer ou de réduire certaines caractéristiques indésirables de la structure d'un échantillon (dissymétrie, variance instable, etc.) en appliquant une fonction mathématique simple à l'ensemble des données. Les "transformations puissances" définies par :

$$x^{(p)} = \begin{cases} . \ x^p \text{ si } p > 0 \\ . \ \log(x) \text{ si } p = 0 \\ . \ -x^p \text{ si } p < 0 \end{cases}$$

sont les fonctions simples les plus couramment utilisées en cas de distributions dissymétriques ou de corrélation entre variance et moyenne. La raison du changement de signe quand $p<0$ est de s'assurer que les valeurs transformées ont le même ordre relatif que les valeurs originales.

Hoaglin et al. (1983) présentent deux méthodes graphiques adaptées à ces problèmes, permettant de déterminer approximativement la valeur de l'exposant p à appliquer : le "spread-versus-level plot" (graphique SVL) et le "transformation plot for symmetry" (graphique S). Ces deux graphiques s'interprètent de manière semblable :

- la linéarité du graphique indique si une transformation puissance est adaptée à l'objectif fixé. Si le graphique n'est pas linéaire, d'autres transformations plus complexes pourront éventuellement être utilisées.

- Si le graphique est à peu près linéaire et de pente b, la transformation puissance appropriée correspond à l'exposant p=1-b (si b=0 alors p=1 et il n'y a pas de transformation à appliquer).

I. LE GRAPHIQUE "SPREAD-VERSUS-LEVEL PLOT".

Lorsqu'on compare plusieurs échantillons, il est fréquent d'observer une relation systématique entre les paramètres de localisation des échantillons (moyenne, médiane, etc.) et leur paramètre de dispersion (écart-type, intervalle interquartile, etc.) (fig.1). Le graphique SVL suggère la transformation puissance à appliquer, si elle existe, pour éliminer ou réduire cette relation systématique.

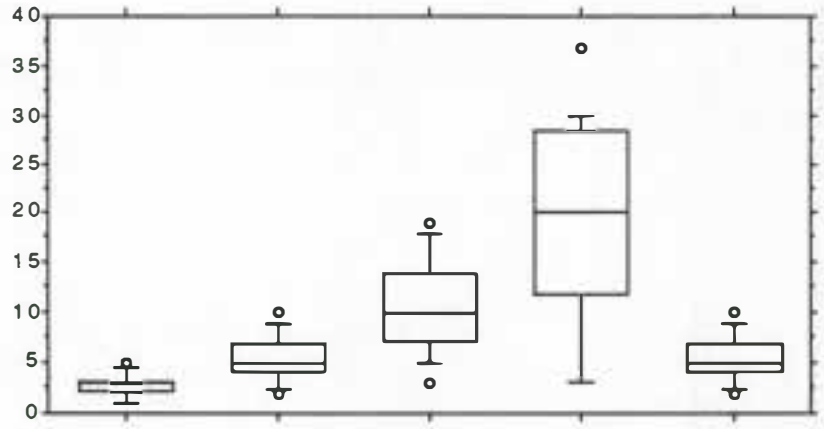


Fig.1 : Box plots de 5 échantillons (la dispersion des échantillons augmente avec leur niveau).

1. Définition du graphique SVL.

Soit K échantillons comparés (k=1,K). Le graphique SVL correspond à la représentation de :

$\log(IQR_k)$ en fonction de $\log(M_k)$
(IQR_k et M_k étant l'intervalle interquartile et la médiane de l'échantillon k).

Si log(IQR)=a+b.log(M) alors la valeur p=1-b est une valeur approximative de l'exposant de la transformation puissance qui stabilise au mieux la dispersion des échantillons étudiés (fig.2).

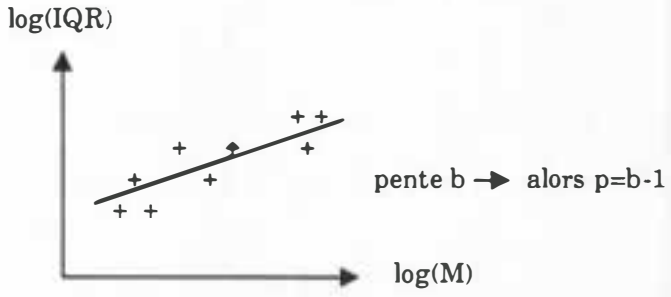


Fig.2 : Principe du graphique SVL ("Spread-Versus-Level plot"). Si la pente de la droite est b, alors l'exposant de la transformation puissance est approximativement p=1-b.

Lors du choix de p, il convient en pratique de préférer une valeur qui correspond à une transformation simple des données. Par exemple, si la valeur p est située entre 0 et 1/2, on

préférera les transformations log ou racine carrée pour des raisons de simplicité et d'interprétabilité des données à des transformations puissances 1/5 ou autres. En fait, la valeur de p est fréquemment arrondie à la plus proche des valeurs du tableau n°1.

Transformation	Power	Slope of Spread-versus-Level Plot
Cube	3	−2
Square	2	−1
No change	1	0 (flat plot)
Square root	½	½
Logarithm	0	1
Reciprocal root	−½	1½
Reciprocal	−1	2

"The transformations in this list are the main members of Tukey's "ladder of powers."

Tab.1 : Transformations puissances les plus utilisées (D'après Hoaglin et al., 1983).

2. Justification du graphique SVL.

Hypothèse : Les variables aléatoires X des différents ensembles de données ont des lois de même forme, mais diffèrent par leur localisation et leur dispersion.

On recherche un exposant puissance p tel que $IQR(X^p)$ soit constant pour ces ensembles de niveaux différents : les dispersions des échantillons transformés doivent être du même ordre.

* Pour $p \neq 0$:

Soit pour X^p : M = médiane
M-c = 1^{er} quartile ($Q_{0.25}$)
M+d = 2^e quartile ($Q_{0.75}$)
d+c = intervalle interquartile (IQR) (d+c = constante par hypothèse)

alors pour X : $M^{1/p}$ = médiane
$(M-c)^{1/p} = Q_{0.25}$
$(M+d)^{1/p} = Q_{0.75}$
$(M+d)^{1/p}-(M-c)^{1/p} = IQR$

On montre pour X (Hoaglin et al., 1983) :

$$IQR = (M+d)^{1/p}-(M-c)^{1/p} = \underbrace{\frac{1}{p}(d+c)M^{(1/p)-1}}_{\text{terme dominant.}}\left[1 + \frac{(1/p)-1}{2}\cdot\frac{(d-c)}{M} + \frac{((1/p)-1)((1/p)-2)}{6}\cdot\frac{d^2+dc+c^2}{M^2} + \ldots.\right]$$

Donc par approximation : $IQR \approx (1/p)(d+c).M^{(1/p)-1}$
et $\log(IQR) \approx \log((1/p)(d+c)) + (1-p)\log(M^{1/p})$.

En résumé : log(IQR de X)≈constante+(1-p)log(médiane de X) et on retrouve la pente b=1-p du graphique SVL. Selon Hoaglin et al. (1983), cette approximation est valable lorsque les quantiles des données transformées ne sont pas trop éloignés de la médiane ($(d/m) \leq 0.4$).

* Pour p = 0 :

Soit pour log(X) : M = médiane (log en base 10)
$M-c = Q_{0.25}$
$M+d = Q_{0.75}$
d+c = IQR (d+c = constante)

alors pour X : 10^M = médiane
$10^{M-c} = Q_{0.25}$
$10^{M+c} = Q_{0.75}$
$10^M(10^d-10^c) = IQR$

et $\log(IQR) = M+\log(10^d-10^c)$
(avec $M = \log(10^M) = \log$(médiane de X))

donc log(IQR de X)=constante+log(médiane de X) (pour p = 0, la pente du graphique SVL est bien b=1).

II. LES GRAPHIQUES DE SYMETRIE.

La symétrie d'un ensemble de données est une propriété souvent recherchée. Elle apporte en général plus de clarté dans la description de cet ensemble, en évitant par exemple l'accumulation et le chevauchement de points dans un angle du graphique. La symétrie facilite aussi la comparaison de plusieurs échantillons ; beaucoup d'estimateurs de paramètres de localisation sont meilleurs et ont plus de sens quand la distribution est symétrique. En outre, il faut noter que les problèmes de dissymétrie (et plus particulièrement de non-normalité) entraînent fréquemment des problèmes de lien entre la localisation et la dispersion des distributions (chap.I).

Une méthode graphique simple pour étudier la symétrie d'une distribution empirique a été proposée par Chambers et al. (1983) :

Soit un échantillon trié par ordre croissant {y(i)} (i=1,n).
La distribution est symétrique si : médiane-y(i)=y(n+1-i)-médiane pour i=1,n
(avec y(i) quantile d'ordre pi=(i-0.5)/n pour tout i).

Il suffit donc de construire le graphique d'abscisses vi=médiane-y(i) et d'ordonnées ui=y(n+1-i)-médiane pour visualiser la symétrie ou la dissymétrie des données. En cas de symétrie, les points (vi,ui) suivent sensiblement la bissectrice u=v.

La figure n^3 présente une distribution qui s'étale nettement vers la droite.

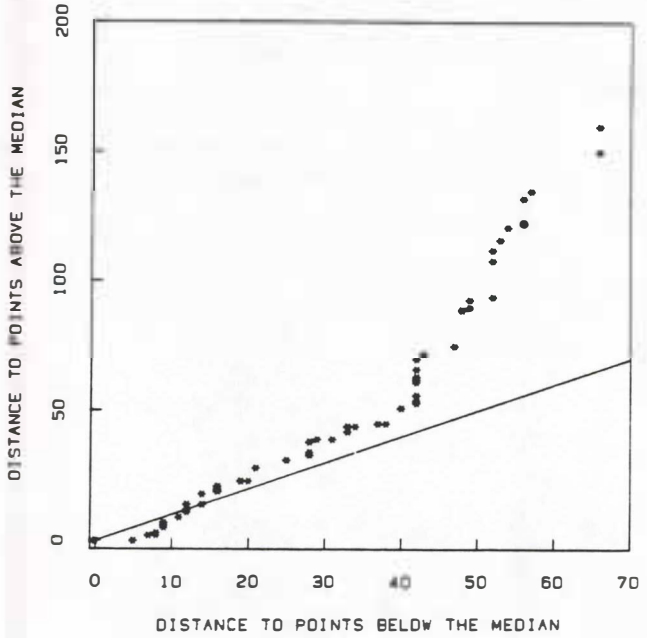


Fig.3 : Représentation d'une distribution dissymétrique à droite. Abscisses : vi=médiane-y(i) ; Ordonnées : ui=y(n+1-i)-médiane. Le trait plein correspond à la bissectrice u=v (D'après Chambers et al., 1983).

Une autre présentation de la condition de symétrie d'un ensemble de données est de définir une valeur moyenne pour toute paire de quantiles symétriques : mid = $(Q_\alpha+Q_{1-\alpha})/2$ (Q_α ,$Q_{1-\alpha}$ étant les quantiles d'ordre α et $1-\alpha$ de la distribution). Si la distribution est symétrique, toutes les valeurs moyennes "mid" sont égales à la médiane (fig.4).

Fig.4 : Valeurs moyennes de quantiles symétriques (mid) dans le cas de distributions symétrique et dissymétrique (M=médiane).

Pour le choix d'une transformation puissance adaptée au problème de dissymétrie, Hoaglin et al. (1983) proposent une méthode graphique : le graphique "S" ("transformation plot for symmetry"), qui utilise cette propriété, en sélectionnant certaines paires de quantiles particuliers : les "letter values" (Tukey, 1977). De la même manière que le graphique SVL, la pente du graphique S renseigne sur l'exposant de la transformation qui symétrise au mieux les données.

1. Les "letter values".

Les "letter values" sont constituées par des paires de quantiles symétriques (Q_α , $Q_{1-\alpha}$) qui débutent avec la médiane, puis les quartiles et enfin qui atteignent les extrêmes en divisant la probabilité a par 2 à chaque fois (tab.2). Chaque lettre correspond à deux valeurs : le quantile inférieur Q_α = Xl et le quantile supérieur $Q_{1-\alpha}$ = Xu. La dénomination des lettres suit la règle suivante : M pour "median", F pour "fourth" (quartile), puis les lettres suivent l'ordre alphabétique inverse.

Tag	Tail Area
M	$\frac{1}{2}$ = .5
F	$\frac{1}{4}$ = .25
E	$\frac{1}{8}$ = .125
D	$\frac{1}{16}$ = .0625
C	$\frac{1}{32}$ = .03125
B	$\frac{1}{64}$ = .015625
A	$\frac{1}{128}$ = .0078125
Z	$\frac{1}{256}$ = .00390625
Y	$\frac{1}{512}$ = .001953125
X	$\frac{1}{1024}$ = .0009765625

Tab.2 : Relation entre les "letter values" et la probabilité α (D'après Hoaglin et al., 1983).

Les auteurs présentent une procédure de calcul des "letter values" très simple :

Soit un échantillon trié {x(i)} i=1,n. On définit la profondeur D de toute valeur x(i) par le plus petit de ses rangs inférieur (= i) et supérieur (=n+i-1). Les valeurs (Xl, Xu) de chaque lettre sont alors déterminées par leur profondeur D :

D = ([profondeur de la lettre précédente]+1)/2 (avec [...] : partie entière)

(à une profondeur D_α correspond deux valeurs : Xl=Q_α et Xu=$Q_{1-\alpha}$).

Cette procédure conventionnelle de calcul des profondeurs correspond à une méthode d'interpolation très simple puisque les "letter values" sont égales soit à l'une des données de l'échantillon, soit à la moyenne de deux données consécutives.

Exemple : Pour l'échantillon trié {x(i)} = {36, 37, 45, 52, 56, 58, 66, 68, 75, 90, 100}, les "letter values" et leur profondeur sont :

"letter values"	Profondeur D	Valeur Xl	Valeur Xu
M	6	58	58
F	3,5	(45+52)/2 = 48.5	(68+75)/2 = 71.5
E	2	37	90
D	1,5	(36+37)/2 = 36.5	(90+100)/2 = 95
.....	.....	.....	.....

2. Définition et justification du graphique S.

Le graphique S se construit d'après le calcul des "letter values" (Xl, Xu) et correspond à la représentation de :

$((Xl+Xu)/2)-M$ en fonction de $((Xu-M)^2+(M-Xl)^2)/4M$
(M : médiane de l'échantillon).

Si le graphique est sensiblement linéaire et de pente b, l'exposant de la transformation puissance à appliquer est approximativement p=1-b.

Justification du graphique S (Hoaglin et al, 1983) :

Soit {xi} i=1,n un échantillon de valeurs positives (s'il existe x<0 , on peut ajouter une constante aux données pour retrouver le cas positif). On cherche p tel que l'échantillon $\{xi^p\}$ i=1,n soit symétrique.

Soit M et (Xl, Xu) la médiane et les "letter values" d'une profondeur fixée. Xl^p, Xu^p et M^p sont alors les quantiles approchés des données transformées d'ordres correspondants.

Pour p≠0, les développements de Taylor de Xl^p et Xu^p à l'ordre 2 en M donnent :

$$Xl^p \approx M^p+pM^{p-1}(Xl-M)+(p(p-1)/2)M^{p-2}(Xl-M)^2$$

$$Xu^p \approx M^p+pM^{p-1}(Xu-M)+(p(p-1)/2)M^{p-2}(Xu-M)^2$$

Si les données transformées sont symétriques alors :

$$((Xl^p+Xu^p)2)-M^p = 0$$

et, en remplaçant par les développements de Taylor, on obtient :

$$((Xl+Xu)/2)-M \approx (1-p)((Xu-M)^2+(M-Xl)^2)/4M.$$

Si la distribution est dissymétrique vers la droite, alors (1-p) sera positif et p<1, et inversement en cas de dissymétrie à gauche.

BIBLIOGRAPHIE.

- CHAMBERS (J.M), CLEVELAND (W.S.), KLEINER (B.), TUKEY (P.A.), 1983. Graphical methods for data analysis. Wadsworth International Group, 395 p.

- HOAGLIN (D.C.), MOSTELLER (F.), TUKEY (J.W.), 1983. Understanding robust and exploratory analysis. Wiley and Sons, Inc., 447 p.

- TUKEY (J.W.), 1977. Exploratory data analysis. Addison Wesley Publishing Company, 506 p.

LES ECHELLES FONCTIONNELLES

Septembre 1992

Matthieu LESNOFF

BIOMETRIE
CIRAD - Forêt

LES ECHELLES FONCTIONNELLES.

I. PRINCIPE DE L'ECHELLE FONCTIONNELLE.

On considère une fonction f monotone croissante $\xi=f(x)$. On appelle échelle fonctionnelle un échelle obtenue en plaçant sur un axe les points x à une distance de l'origine x_0 égale à $\xi-\xi_0$ (fig.1).

Fig.1 : Echelle fonctionnelle (D'après Callot, 1984).

La figure n°2 présente un exemple d'échelle logarithmique :

Fig.2 : Echelle fonctionnelle logarithmique (D'après Callot, 1984).

On appelle graphique fonctionnel un système d'axes orthogonaux à échelles fonctionnelles $\xi=f(x)$ et $\eta=g(y)$. Si l'on considère une courbe C d'équation $y=\varphi(x)$ sur papier arithmétique, l'image de φ sur le graphique (ξ,η) est $\varphi^*=g[\varphi(f^{-1})]$.

La pente de la tangente à la courbe est : $d\eta/d\xi=g'(y)dy/f'(x)dx$. En plus de cette pente, deux éléments différentiels sont spécialement intéressants :

- le taux d'accroissement (accroissement relatif) : $(1/y)(dy/dx)$.

- le quotient des variations relatives (élasticité) : $(dy/y)/(dx/x)$.

Il est possible de rechercher les graphiques fonctionnels qui permettent de représenter ces éléments différentiels par la pente $d\eta/d\xi$ de la tangente à la courbe $\eta=\varphi^*(\xi)$; l'intérêt est que celle-ci peut être appréciée facilement de manière visuelle sur le graphique.

On montre (Callot, 1984) :

- $g'(y)dy/f'(x)dx=K.(1/y)(dy/dx)$ ssi $f(x)=ax+b$ et $g(y)=c\ln(y)+d$

La pente de la tangente est donc proportionnelle au taux d'accroissement si on construit un graphique semi-logarithmique. Ce graphique se définit par une échelle arithmétique en abscisse et logarithmique en ordonnée.

- $g'(y)dy/f'(x)dx=K.(dy/y)/(dx/x)$ ssi $f(x)=a\ln(x)+b$ et $g(y)=c\ln(y)+d$

La pente de la tangente est donc proportionnelle au quotient des variations relatives si l'on construit un graphique logarithmique. Ce graphique se définit par deux échelles logarithmiques en abscisse et en ordonnée.

Il faut noter que l'utilisation de logarithmes de bases différentes ne change pas la forme des graphiques semi-logarithmiques et logarithmiques, mais influence uniquement la valeur des axes. Les logarithmes de bases différentes sont multiples les uns des autres : $\log_b(x)=\log_a(x)/\log_a(b)$. Le choix de la base dépend de l'étendue des valeurs étudiées. Cleveland (1985) conseille d'utiliser le logarithme de base 2 pour des valeurs allant jusqu'à 10^3 ($2^8=1024$). Le logarithme de base 10 est bien adapté lorsque les données s'étendent sur une grande gamme de puissances de 10.

II. UTILISATION DES GRAPHIQUES FONCTIONNELS.

1. Les graphiques semi-logarithmiques.

1.1 Etude des taux d'accroissement.

La propriété principale du graphique semi-logarithmique est que les pentes des courbes (ou des segments de courbes) qu'il représente renseignent directement sur les variations des taux d'accroissement des variables considérées, que ces taux soient positifs ou négatifs (fig.3).

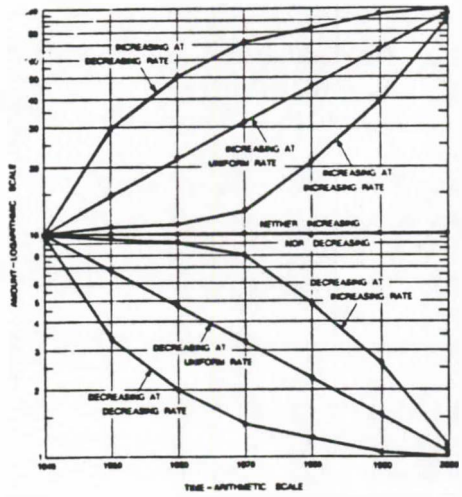


Fig.3 : Schéma classique de courbes sur un graphique semi-logarithmique (D'après Schmidt, 1986).

Le taux d'accroissement entre deux points (x_1,y_1) et (x_2,y_2) est défini par :

$$t=(\ln(y_2)-\ln(y_1))/(x_2-x_1)$$

Une fois calculés sur les segments de courbes désirés, ces taux peuvent être reportés le long des courbes du graphique (fig.4). Ceci facilite l'appréciation de leur évolution sur une même courbe, ainsi que leur comparaison entre les différentes courbes.

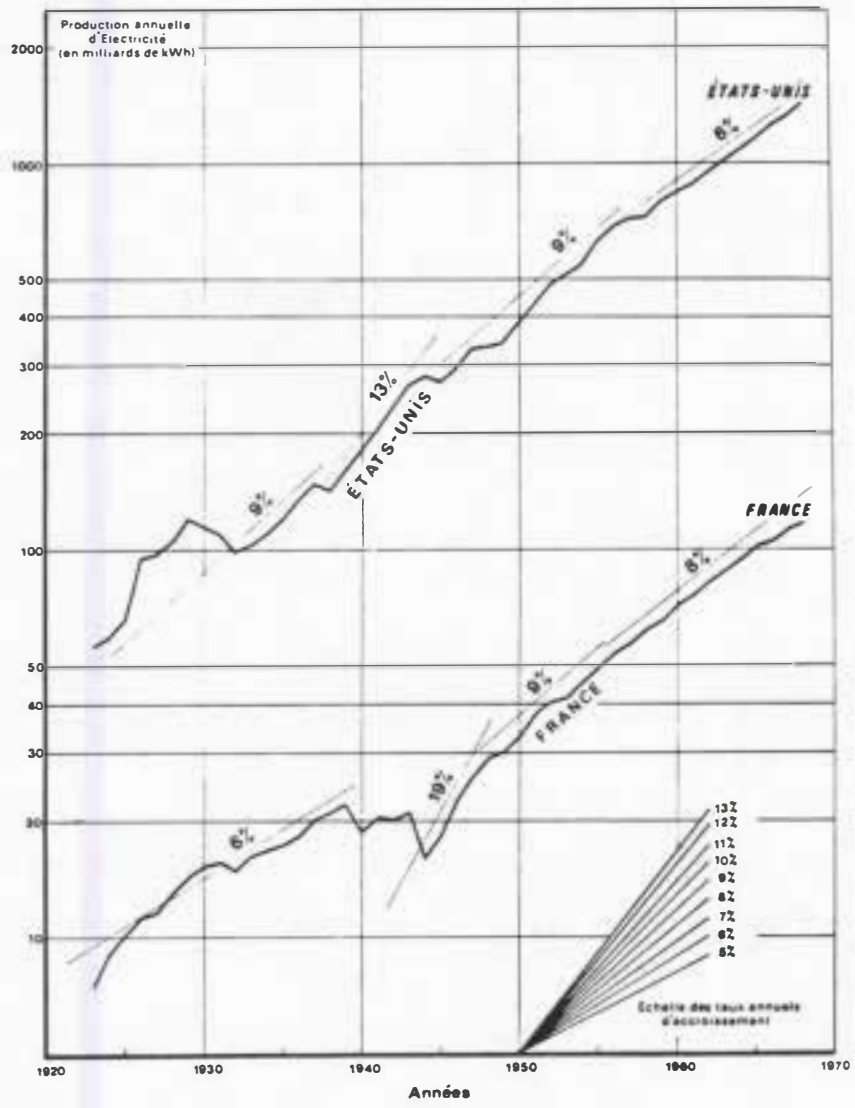


Fig.4 : Graphique semi-logarithmique de la production annuelle d'électricité aux Etats-Unis et en France de 1923 à 1968. Les taux d'accroissement ont été reportés le long des courbes (D'après Callot, 1984).

L'utilisation de l'échelle logarithmique permet souvent d'améliorer la lisibilité d'un graphique, en particulier lorsqu'on cherche à comparer des courbes qui varient dans des gammes de valeurs très différentes. L'échelle logarithmique fournit en effet une précision relative constante en dilatant les zones correspondant aux valeurs les plus faibles et en réduisant celles correspondant aux valeurs les plus fortes. Schmidt (1986) prend l'exemple de l'étude comparée des importations en différentes viandes rouges aux Etats-Unis entre 1960 et 1975 (fig.5).

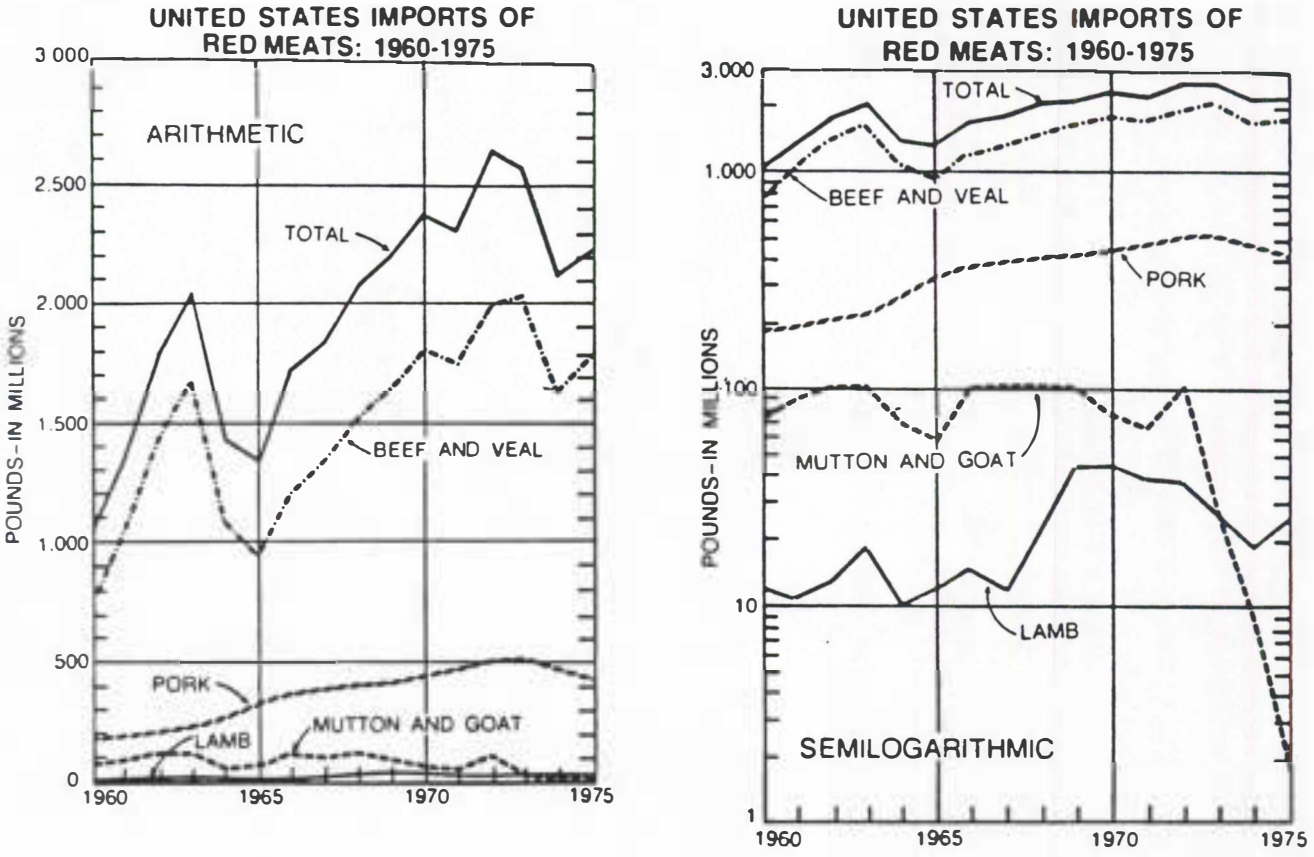


Fig.5 : Importations en différentes viandes rouges aux Etats-Unis entre 1960 et 1975 (gauche : graphique arithmétique ; droite: graphique semi-logarithmique) (D'après Schmidt, 1986).

Du fait de la dominance des importations en viande de boeuf et de veau, le graphique arithmétique simple (à gauche) ne permet pas d'apprécier convenablement les variations des courbes d'importations des autres produits. Il est clair que l'utilisation d'une échelle logarithmique (à droite) permet de visualiser beaucoup mieux les variations de l'ensemble des courbes représentées. Les importations en viande de mouton et de chèvre et en viande d'agneau montrent ainsi des variations importantes qui n'apparaissaient pas sur le premier graphique.

1.2 Linéarisation de courbes.

La courbe représentative de la fonction y(x) est une droite sur un graphique semi-logarithmique si le taux d'accroissement t est constant : t=(1/y)(dy/dx)=constante.

D'où : dy/y=tdx et par intégration : $y=y_0\exp\{t(x-x_0)\}$.

Ainsi, les courbes exponentielles deviennent des droites sur un graphique semi-logarithmique.

2. Le graphique logarithmique.

2.1 Etude du quotient des variations relatives (élasticité).

L'élasticité, grandeur très utilisée en économie, correspond au quotient des variations relatives (dy/y)/(dx/x) calculé sur le segment de courbe étudié. Entre deux points (x_1,y_1) et (x_2,y_2), elle est définie par :

$$e=(\ln(y_2)-\ln(y_1))/(\ln(x_2)-\ln(x_1)).$$

La figure n°6 montre un exemple d'utilisation du graphique logarithmique (Callot, 1984) en représentant la répartition des dépenses des ménages non agricoles en 1956 :

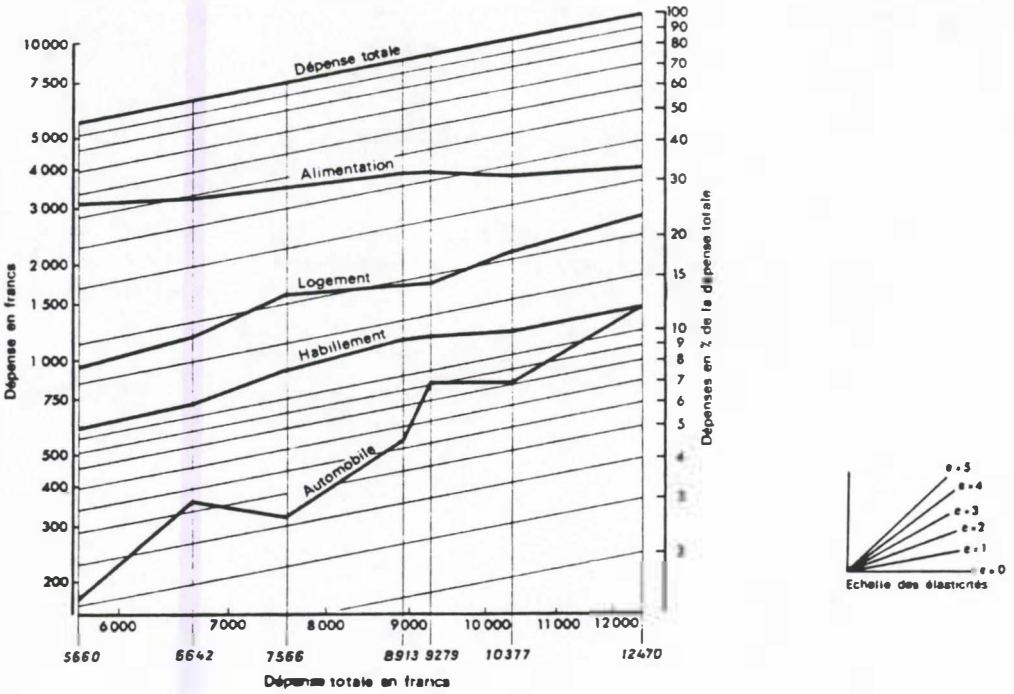


Fig.6 : Répartition des dépenses des ménages non agricoles en 1956 (D'après Callot, 1984).

Les quatre dépenses partielles figurant sur le graphique présentent des élasticités pratiquement constantes par rapport à la dépense totale (les courbes sont approximativement linéaires) :

alimentation : e=0.4
logement : e=1.3
habillement : e=1.1
automobile : e=2.6

Pour un taux d'accroissement donné de la dépense totale, le taux d'accroissement de la dépense en alimentation est 0.4 fois plus faible que celui-ci (dépense peu élastique), et celui de la dépense en automobile est 2.6 fois plus fort (dépense très élastique).

Autrement dit, deux ménages dont les dépenses totales diffèrent de 10% ont des dépenses en alimentation qui diffèrent de 4% et des dépenses en automobile qui diffèrent de 26%.

2.2 Linéarisation de courbes.

La courbe représentative de la fonction y(x) est une droite sur un graphique logarithmique si l'élasticité e est constante : e=(dy/y)/(dx/x)=constante.

D'où : $dy/y=edx/x$ et par intégration : $y=ax^e$.

Ainsi, les fonctions puissances deviennent des droites sur un graphique logarithmique. Cette propriété est souvent utilisée lorsqu'on veut linéariser des courbes de croissance allométrique de la forme $y=ax^b$, et en estimer les paramètres :

$y=ax^b=a.\exp\{b\ln(x)\}$ d'où : $\ln(y)=\ln(a)+b\ln(x)$

ln(a) et b sont estimés par régression linéaire de ln(y) en fonction de ln(x). On en déduit l'estimateur (biaisé) de a : $\exp\{\widehat{\ln}(a)\}$.

Une autre propriété importante du graphique logarithmique est que, d'une part, les hyperboles xy=constante sont transformées en droites parallèles à la deuxième bissectrice (fig.7), et d'autre part, que les droites y/x=constante sont transformées en droites parallèles à la première bissectrice (fig.8).

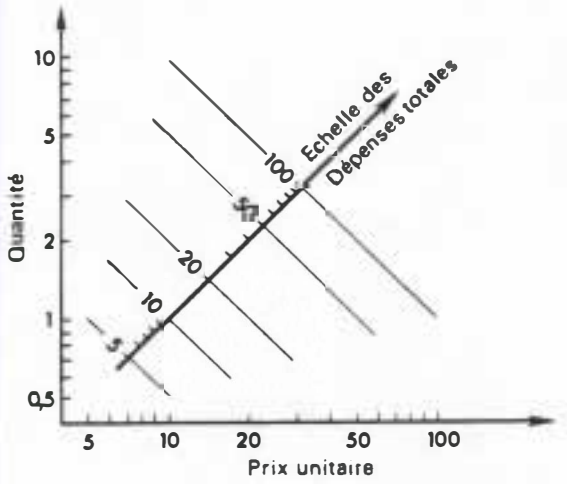


Fig.7 : Graphique logarithmique : les hyperboles xy=constante sont des droites parallèles à la deuxième bissectrice (D'après Callot, 1984).

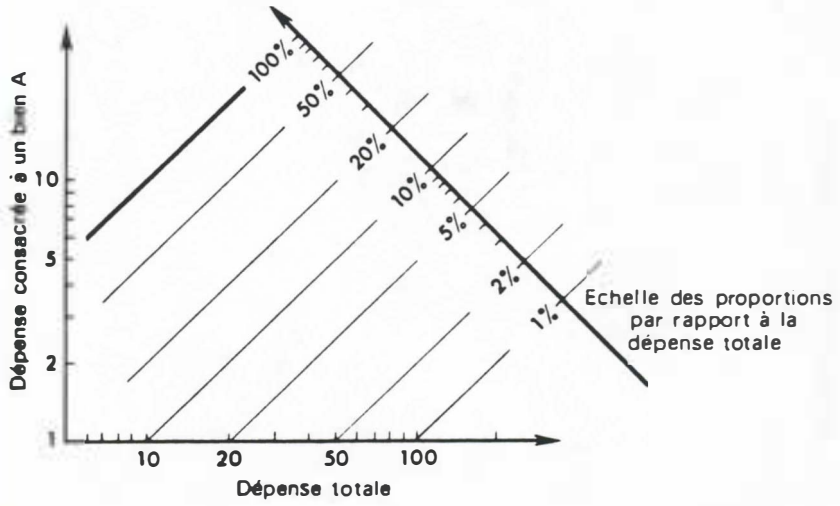


Fig.8 : Graphique logarithmique : les droites x/y=constante sont des droites parallèles à la première bissectrice (D'après Callot, 1984).

BIBLIOGRAPHIE.

- CALLOT (G.), 1984. Cours de statistique descriptive. Dunod, 488 p.

- SCHMIDT (C.F.), 1986. Whatever has happened to the semilogarithmic chart. The American Statistician, 40(3): 239-244.

REPRESENTATION D'UN TABLEAU CROISE PAR LE GRAPHIQUE "TWO-WAY PLOT" DE J.W. TUKEY

Septembre 1992

Matthieu LESNOFF

BIOMETRIE
CIRAD - Forêt

REPRESENTATION D'UN TABLEAU CROISE PAR LE GRAPHIQUE 'TWO-WAY PLOT' DE J.W. TUKEY.

Tukey (1977) a proposé une méthode graphique pour représenter un tableau croisé, le "two-way plot", qui permet d'étudier l'additivité des deux facteurs du tableau considéré :

Soit deux facteurs A et B croisés en un tableau Y(i,j). On peut construire le modèle additif : $y_{ij}=m+\alpha_i+\beta_j+e_{ij}$ (avec e_{ij} : résidus). Les estimations $\hat{y}_{ij}=\hat{m}+\hat{\alpha}_i+\hat{\beta}_j$ sont obtenues par les méthodes d'analyse de variance classiques.

Le two-way plot est alors la représentation de:

$$\hat{y}_{ij} \text{ en fonction de } h_{ij}=\hat{\alpha}_i-\hat{\beta}_j.$$

Le graphique a la forme d'une grille penchée dont chaque point représente le croisement des modalités (i,j) du tableau de données. On lit la valeur de $\hat{y}_{ij}$ sur l'échelle verticale. Les coordonnées h_{ij} n'ont pas d'interprétation. Les résidus $\hat{e}_{ij}$ du modèle additif peuvent être représentés sur la grille du two-way plot par différents symboles selon leur valeur (il faut préalablement constituer des classes de résidus).

Le two-way plot de la figure n°1 présente le tableau des températures mensuelles moyennes de trois villes américaines (tab.1). Sept classes de résidus ont été constituées et représentées symboliquement sur la grille du graphique. On peut remarquer que le two-way plot permet de repérer facilement les modalités (i,j) fortes et les modalités (i,j) faibles, ce qui n'était pas évident à partir du tableau lui-même.

	Flagstaff	Phoenix	Yuma
July	65.2	90.1	94.6
Aug	63.4	88.3	93.7
Sept	57.0	82.7	88.3
Oct	46.1	70.8	76.4
Nov	35.8	58.4	64.2
Dec	28.4	52.1	57.1
Jan	25.3	49.7	55.3

Tab.1 : Températures mensuelles moyennes (°F) des villes de Flagstaff, Phoenix et Yuma (D'après Tukey, 1977).

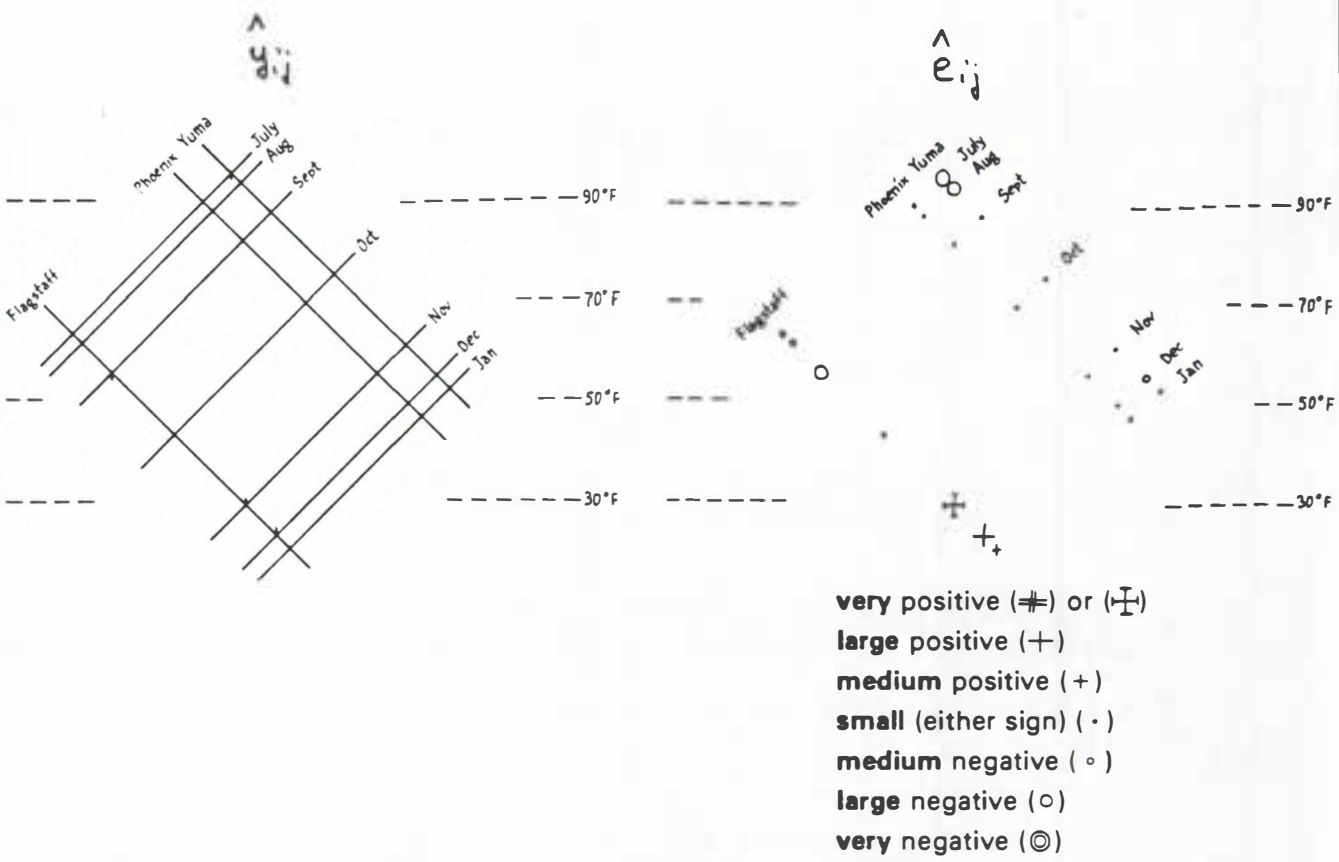


Fig.1 : Two-way plot du tableau n°1. Les estimations $\hat{y}_{ij}$ correspondent au modèle additif (D'après Tukey, 1977).

La représentation symbolique des résidus est très utile pour repérer certaines déviations par rapport à l'additivité des facteurs. Par exemple, sur la figure n°2, les tendances opposées des résidus pour les modalités 1 et 3 du facteur A suggèrent l'existence d'une interaction AxB.

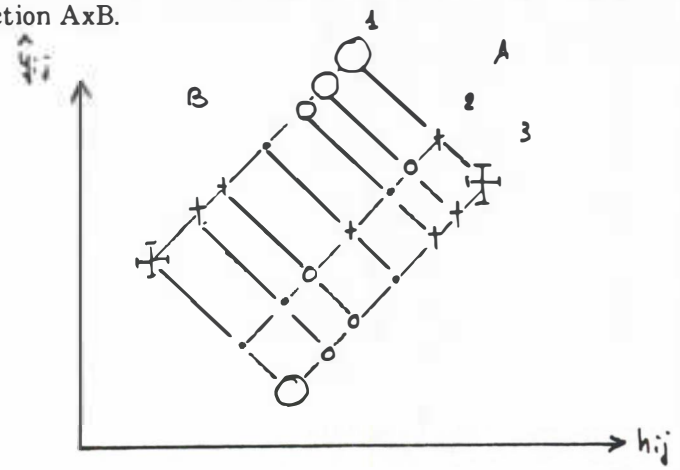


Fig.2 : Two-way plot d'un tableau croisé non additif avec représentation symbolique des résidus.

En agronomie, le two-way plot s'adapte très bien aux plans en blocs aléatoires complets. Chambers et al. présentent l'exemple de l'analyse des rendements en orge dans quatre localités (blocs) du Minnesota (tab.2). Les résidus du modèle additif sont représentés par des flèches dirigées vers le haut ou vers le bas selon leur signe, et de hauteur proportionnelle à leur valeur (fig.3). Pour conserver une certaine lisibilité, seuls les résidus extrêmes sont considérés. L'opposition des tendances résiduelles des deux meilleures variétés (opposition des flèches) suggère l'existence d'une interaction bloc-variété.

	Station					
Variety	**1**	**2**	**3**	**4**	**5**	**6**
Manchuria	162	247	185	219	165	155
Svansota	187	258	182	183	139	144
Velvet	200	263	195	220	166	146
Trebi	197	340	271	267	151	194
Peatland	182	254	220	201	184	190

Tab.2 : Rendements de 6 variétés d'orge dans 5 localités du Minnesota entre 1930-1931 (D'après Chambers et al., 1983).

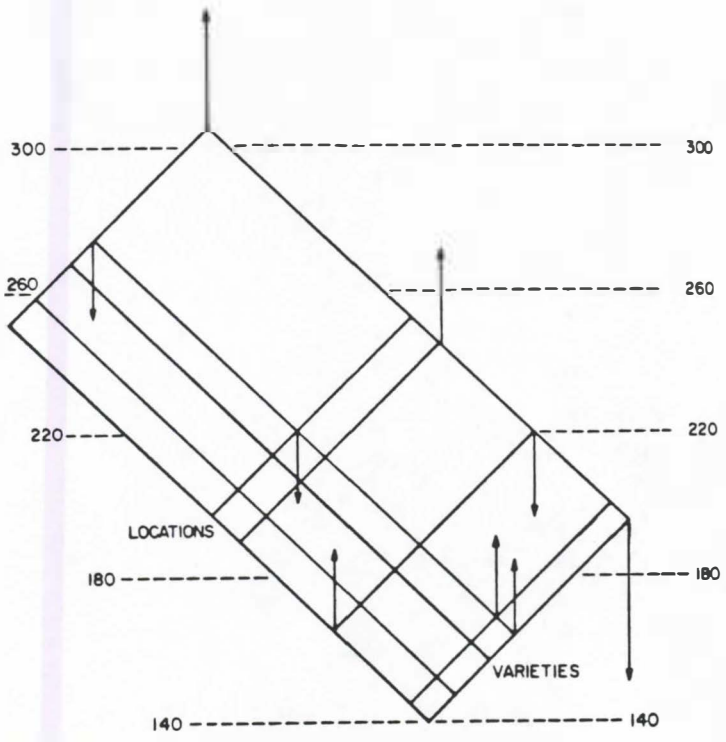


Fig.3 : Two-way plot du tableau n°2. Les flèches représentent les résidus extrêmes (D'après Chambers et al., 1983).

BIBLIOGRAPHIE.

- CHAMBERS (J.M), CLEVELAND (W.S.), KLEINER (B.), TUKEY (P.A.), 1983. Graphical methods for data analysis. Wadsworth International Group, 395 p.

- TUKEY (J.W.), 1977. Exploratory data analysis. Addison Wesley Publishing Company, 506 p.

PROBLEMES GRAPHIQUES DE COMPARAISONS DE COURBES ET DE NUAGES DE POINTS

Septembre 1992

Matthieu LESNOFF

BIOMETRIE
CIRAD - Forêt

PROBLEMES GRAPHIQUES DE COMPARAISONS DE COURBES ET DE NUAGES DE POINTS.

1. Ecart entre deux courbes.

Il est très courant de superposer deux courbes sur le même graphique pour essayer d'en apprécier les écarts. Cleveland (1985) donne l'exemple du statisticien W. Playfair (1786) qui cherchait à comparer les importations et les exportations des *East Indies* entre 1700 et 1780 (fig.1). Il est possible de mesurer l'écart entre ces deux courbes par la différence entre celles-ci point par point (c'est-à-dire par des écarts absolus). Le problème est qu'il est très difficile pour notre système visuel d'apprécier des variations de distances verticales en particulier quand les courbes ont des pentes variables. Par exemple, d'après la fig.1, l'impression visuelle est, d'une part, qu'il y a une faible différence entre exportations et importations entre 1755 et 1765, et d'autre part, que cette différence est constante. Hors, la fig.3 montre qu'il n'en est rien : on observe un pic dans la courbe "différence" juste après 1760.

Un autre exemple met en évidence cette difficulté d'appréciation visuelle : les deux courbes exponentielles de la fig.2 paraissent se rapprocher plus on s'éloigne de l'origine. En fait, leur écart absolu reste constant sur l'ensemble du graphique (fig.4).

Notre système visuel est donc en général inadapté pour juger des écarts entre deux courbes. Le graphique des différences point par point apparaît très utile pour mieux apprécier les variations de distances verticales. Ce même type de graphique peut d'ailleurs être appliqué lorsqu'il s'agit de comparer un nuage de points à une courbe (graphique des "résidus" : différences entre les points du nuage et la courbe).

2. Discrimination de courbes et de nuages de points.

Lorsqu'on cherche à comparer plusieurs groupes d'individus (des espèces par exemple) mesurés sur deux variables communes X et Y, Cleveland et McGill (1984 b) et Cleveland (1985) soulignent l'importance du choix des modes de représentation graphique dans l'appréciation du changement de la relation $y = f(x)$ en fonction des groupes considérés.Deux options se présentent pour les graphiques bivariés.

La première est de superposer les représentations (courbes ou nuages) de tous les groupes sur le même graphique. Si la discrimination des différents groupes est satisfaisante, cette option est assez performante puisqu'elle permet une comparaison directe de leurs caractéristiques (minima, maxima, pentes, etc.).

La deuxième option, qui peut s'appliquer lorsque la superposition des groupes donne un graphique peu lisible, est d'éclater cette représentation en autant de graphiques que de groupes étudiés, et de les juxtaposer. La discrimination des groupes est alors maximale mais la comparaison des positions relatives de chacun des ensembles devient plus difficile.

Cleveland (1985) réduit ce problème en introduisant des repères visuels dans chaque graphique.

2.1 La superposition.

La superposition de plusieurs courbes ou de plusieurs nuages de points sur un même graphique pose le problème de leur identification visuelle. Si deux courbes ont des pentes très différentes à leur point d'intersection, il est relativement aisé de les différencier rapidement (ex : courbes 3 et 5 de la fig.5). Mais si leur pente est similaire, leur identification devient beaucoup plus fastidieuse, ce qui nuit à l'interprétation du graphique (ex : courbes 3 et 2 de la fig.5).

La discrimination des courbes superposées peut être améliorée en utilisant différents motifs pour chaque courbe (fig.6), mais surtout par l'utilisation de la couleur (fig.9) qui augmente considérablement le pouvoir de discrimination visuelle.

En ce qui concerne les nuages de points, Cleveland et McGill (1984 b) et Lewandowsky et Spence (1989) ont proposé un classement global des techniques graphiques, la couleur fournissant la meilleure discrimination, suivie par les motifs plus ou moins remplis puis les motifs de différentes formes, et enfin par les motifs "lettres" (fig.7). Cleveland (1985) fournit deux ensembles de motifs, qu'il suggère d'utiliser dans différentes conditions (fig.8) : l'ensemble supérieur est à préférer lorsqu'il n'y a pas trop de chevauchements de points, l'ensemble inférieur lorsque ce problème intervient. Les deux premiers motifs de chaque ensemble sont à utiliser quand il y a deux groupes à comparer, les trois premiers quand il y a trois groupes, etc.

Si les motifs "lettres" sont parfois critiqués pour leur faible pouvoir de discrimination, ce pouvoir dépend en fait beaucoup du choix des lettres utilisées pour représenter les groupes : les lettres E, F, H rendent difficile l'identification des groupes alors que les lettres H, Q, X sont plus performantes (Lewandowsky et Spence, 1989). En outre, les lettres ont parfois l'avantage de permettre d'identifier les groupes directement sans avoir besoin de légende : M et F pour des populations mâle et femelle par exemple. Elles peuvent alors devenir meilleures que les motifs plus ou moins remplis ou les motifs de différentes formes.

En résumé, l'ensemble des auteurs s'accordent sur l'intérêt de la couleur qui fournit une discrimination excellente pour les différents groupes codés. Pour optimiser cette discrimination, il convient de choisir des couleurs qui s'assemblent mal et qui choquent (éviter les graphiques esthétiques). Deux exemples sont présentés aux fig.9 et fig.10.

2.2 La juxtaposition.

Dès que le nombre d'éléments représentés est important, il survient fréquemment des problèmes de chevauchements de points ou de courbes qui peuvent rendre le graphique illisible. Une alternative à la technique de superposition est alors de construire autant de graphiques juxtaposés que d'éléments représentés.

Cette technique nécessite, pour être efficace, de placer des repères visuels identiques dans chaque graphique, qui permettent de juger des positions relatives des différents ensembles. Ces repères peuvent être des droites particulières dans le cas de juxtapositions de courbes (fig.11), ou des droites de régression et des courbes lissées dans le cas de juxtapositions de nuages de points (fig.12 et fig.13).

3. Représentations schématiques de nuages de points.

3.1 Le lissage polaire et les ellipses de dispersion.

Il est possible de remplacer un ensemble de nuages de points par des "patates" ou des ellipses superposées, qui permettent une vision rapide des principales caractéristiques des nuages dans le plan (X,Y) : Comment se positionnent-ils globalement les uns par rapport aux autres ? Ont-ils des directions particulières (études des variances σ_x et σ_y)? Quelle est l'orientation de ces directions (étude des corrélations $\rho(X,Y)$) ? Ces "résumés" de nuages peuvent être très utiles lorsqu'un trop grand nombre de point empêche les méthodes de représentation classiques, et sont ainsi une alternative à la méthode de juxtaposition présentée ci-dessus.

Cleveland (1984) a introduit une méthode de lissage polaire pour obtenir une telle représentation schématique des nuages (fig.14) :

> Pour chaque nuage, les points M_i de coordonnées cartésiennes (x_i,y_i) sont centrés et réduits, et subissent une rotation qui éliminent la corrélation entre les variables X et Y. Les nouvelles coordonnées (η_i,ζ_i) sont ensuite transformées en coordonnées polaires (r_i,θ_i), l'origine étant placé au centre du graphique. Les coordonnées ri sont alors lissées en fonction des coordonnées θi. Les coordonnées (r_i^*,θ_i) sont alors retransformées en coordonnées cartésiennes "lissées" (x_i^*,y_i^*) dans le repère original du graphique. Il suffit ensuite de joindre les points (x_i^*,y_i^*) dans l'ordre des θ_i croissants, la dernière étape étant la réunion des points (x_1^*,y_1^*) et (x_n^*,y_n^*), pour obtenir les formes ellipsoïdales de la figure n°14.

Une idée similaire est de remplacer chacun des nuages étudiés par son ellipse de dispersion :

> Si le vecteur aléatoire $(X\ Y)'$ suit une loi normale bidimensionnelle, sa fonction de densité f définit dans l'espace (X,Y,f) une surface en forme de cloche dont le sommet est situé au niveau du point moyen (μ_x,μ_y) (fig.15). Les intersections de cette surface avec des plans horizontaux correspondent à des lignes d'égale densité et constituent des ellipses (ou des cercles si $\sigma_x=\sigma_y$) appelée ellipses de dispersion. Chaque ellipse E_p se définit par rapport à une probabilité p d'inclusion d'une valeur de (X,Y) dans celle-ci : $P((X,Y) \subset E_p)=p$.
>
> L'orientation d'une ellipse de dispersion dépend directement du coefficient de corrélation $\rho(X,Y)$. Si les deux variables sont indépendantes, les axes de symétrie de l'ellipse sont parallèles aux axes du repère (X,Y).
>
> L'équation d'une ellipse de dispersion peut se calculer d'après le théorème suivant (Mardia et al., 1979) :
>
> > Si le vecteur aléatoire Z_n à n dimensions suit une loi normale $N_n(\mu,\Sigma)$, alors $(Z_n-\mu)'\Sigma^{-1}(Z_n-\mu)$ suit une loi du Chi-deux à n d.d.l. .
>
> Appliqué au vecteur $(X\ Y)'$, ce théorème devient :
>
> $$((X-\mu_x)\ (Y-\mu_y))\Sigma^{-1}((X-\mu_x)\ (Y-\mu_y))' \approx \chi^2_{(2)}$$

ce qui permet d'écrire, en choisissant une probabilité d'inclusion $p=P(\chi^2_{(2)} \leq Q_\chi(p))=P((X,Y) \in E_p)$, l'équation de l'ellipse E_p :

$$\frac{1}{\sigma_x^2\sigma_y^2-\sigma_{xy}^2}[(\sigma_y^2(x-\mu_x)^2-2\sigma_{xy}(x-\mu_x)(y-\mu_y)+\sigma_x^2(y-\mu_y)^2]=Q\chi(p)$$

(avec $Q_\chi(p)$ le quantile d'ordre p du $\chi^2_{(2)}$).

En supposant que les deux variables ne s'éloignent pas trop de la normalité, il suffit donc d'estimer leur moyenne, leur variance et leur covariance pour chacun des nuages, de choisir une probabilité d'inclusion (ex : p=0.95), puis de tracer les différentes ellipses à l'aide de l'équation présentée ci-dessus. Chaque ellipse contiendra alors environ 100p% du nuage qu'elle représente. Sur le graphique, il est possible de conserver les points extérieurs aux ellipses pour chaque nuage (points extrêmes). Cinq ellipses de dispersion $E_{0.95}$ ont été tracées pour les nuages de la figure n°16, qui correspondent à des échantillons de bois de densités différentes (fig.17).

3.2 Les box plots à deux dimensions.

Une autre approche de représentation schématique est de construire un *box plot* bidimensionnelle pour chaque nuage étudié. La figure n°19 reprend l'exemple des échantillons de bois de densités différentes. De la même manière que les courbes lissées ou les ellipses de dispersion, les *box plots* à deux dimensions sont avantageux lorsqu'un trop grand nombre de points rend illisible les représentations classiques (fig.16 et fig.18). Ils sont en outre très facile à mettre en oeuvre puisqu'il suffit de calculer la médiane et les quartiles des variables X et Y pour chacun des nuages. Dans le repère (X,Y), la hauteur du rectangle représente l'intervalle interquartile de Y et sa largeur l'intervalle interquartile de X. Les barres intérieures verticale et horizontale sont centrées respectivement sur les médianes de X et de Y .

Par souci de lisibilité, les bras des *box plot*s de la figure n°19 n'ont pas été représentés. Il est par contre intéressant de conserver les points extrêmes (*outliers*) de chaque nuage, qui sortent des intervalles $[Q_x(0.25)-1.5IQR_x ; Q_x(0.75)+1.5IQR_x]$ et $[Q_y(0.25)-1.5IQR_y ; Q_y(0.75)+1.5IQR_y]$.

Il faut enfin noter que les *box plots* bidimensionnels correspondent à une représentation orthogonale des variables X et Y. Ils ne tiennent donc pas compte des corrélations entre ces deux variables, à la différence des ellipses de dispersion.

BIBLIOGRAPHIE.

- CLEVELAND (W.S.), McGILL (R.), 1984 a. Graphical perception : theory, experimentation, and application to the development of graphical methods. JASA, 79(387) : 531-554.

- CLEVELAND (W.S.), McGILL (R.), 1984 b. The many faces of a scatterplot. JASA, 79(388) : 807-822.

- CLEVELAND (W.S.), 1985. The element of graphical data. Wadsworth Advanced Books and Software, 329 p.

- LEWANDOWSKY (S.), SPENCE (J.), 1989. Discriminating strata in scatterplots. JASA, 84(407) : 682-688.

- MARDIA (K.V.), KENT (J.T), BIBBY (J.M.), 1979. Multivariate analysis. Academic Press, London, 521 p.

- PLAYFAIR (W.), 1786. The commercial and political atlas. London.

- VENTSEL (H.), 1973. Théorie des probabilités. Mir, Moscou, 563 p.

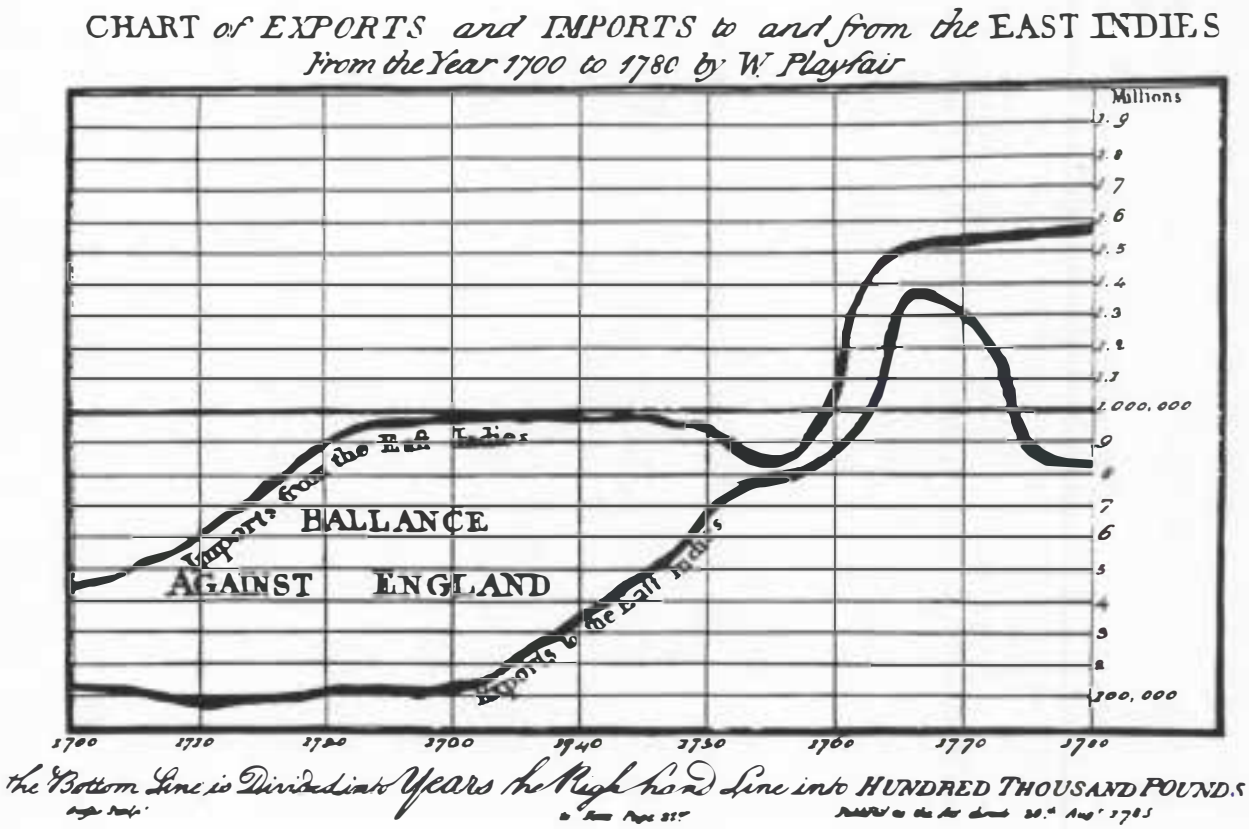


Fig.1 : Courbes d'exportations et d'importations des "East Indies" entre 1700 et 1780 (D'après Playfair, 1786 dans Cleveland, 1985).

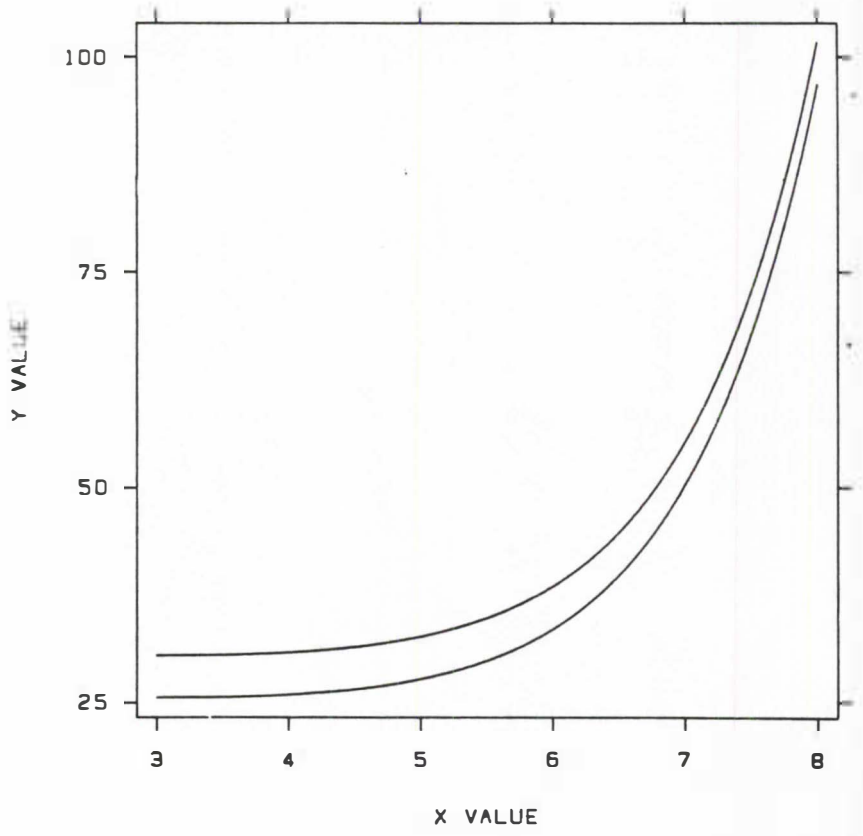


Fig.2 : Comparaison de deux courbes exponentielles (D'après Cleveland, 1985).

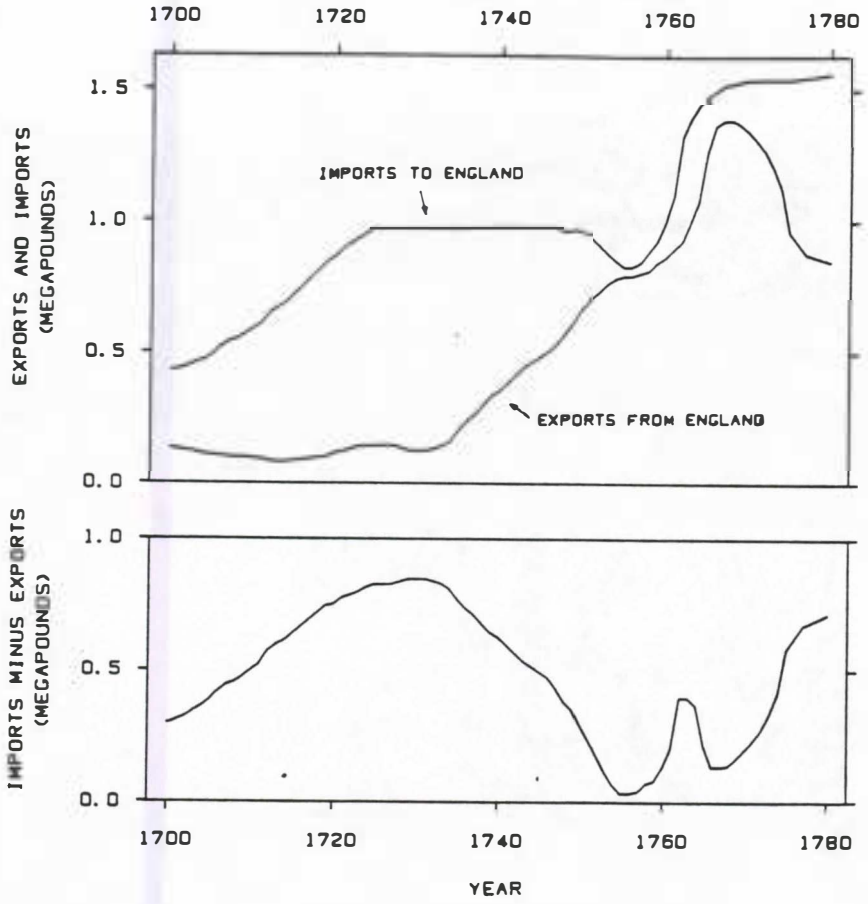


Fig.3 : Courbes d'exportations et d'importations (W. Playfair) (haut) et courbe différence (bas) (D'après Cleveland, 1985).

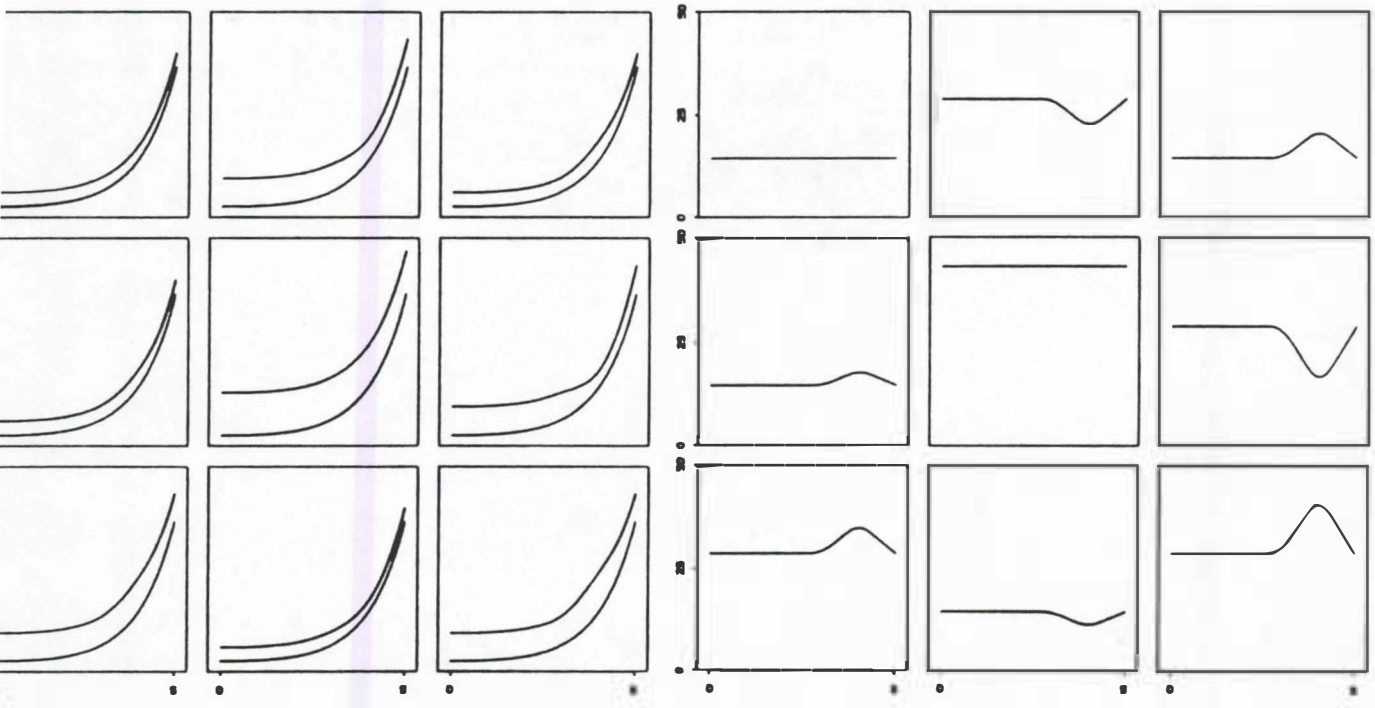


Fig.4 : Comparaisons de courbes exponentielles (gauche) et courbes "différences" (D'après Cleveland et McGill, 1984 a).

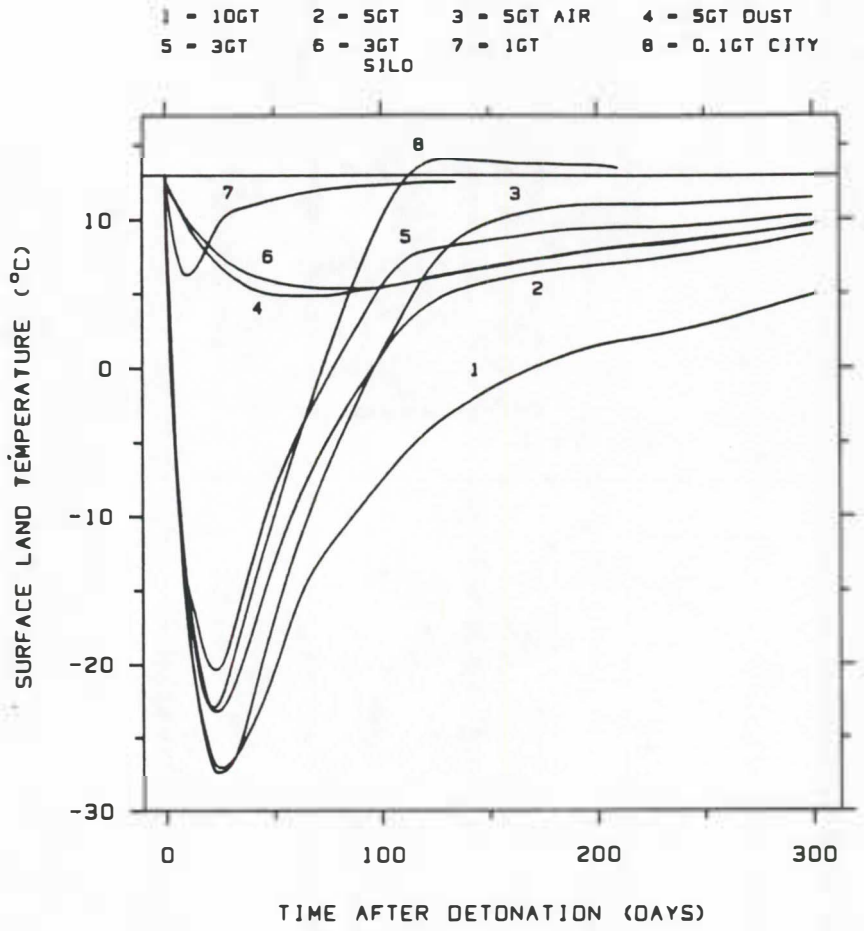


Fig.5 : Superposition de courbes sans motifs (D'après Cleveland, 1985).

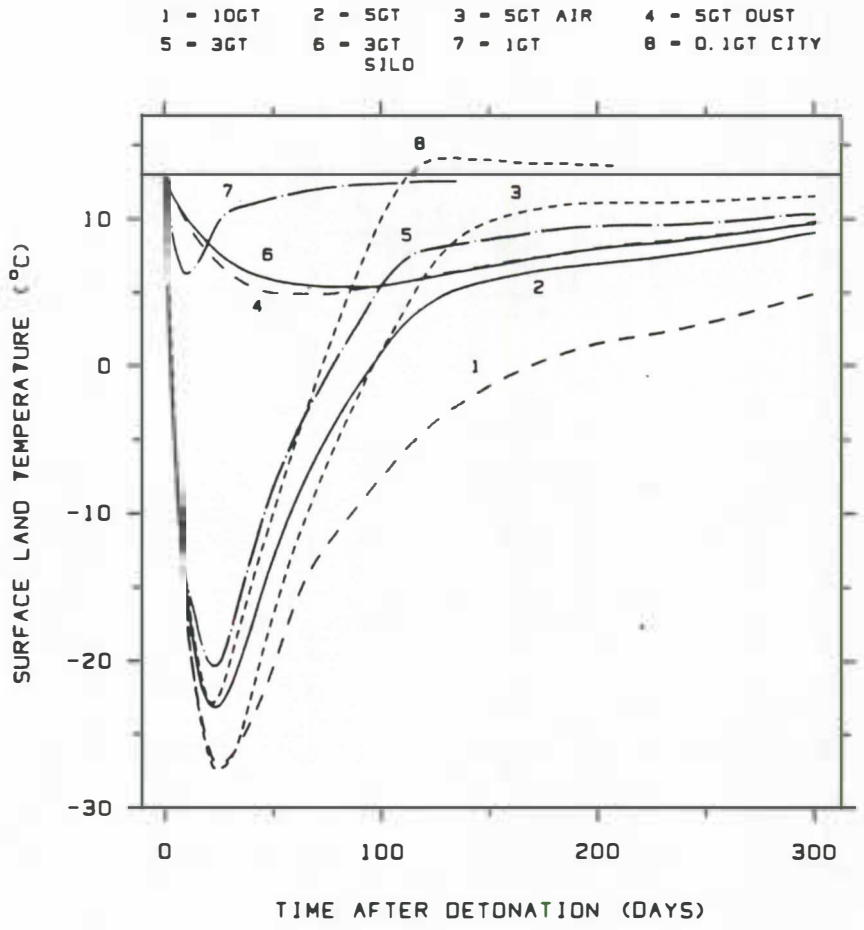


Fig.6 : Superposition de courbes ayant des motifs différents (D'après Cleveland, 1985).

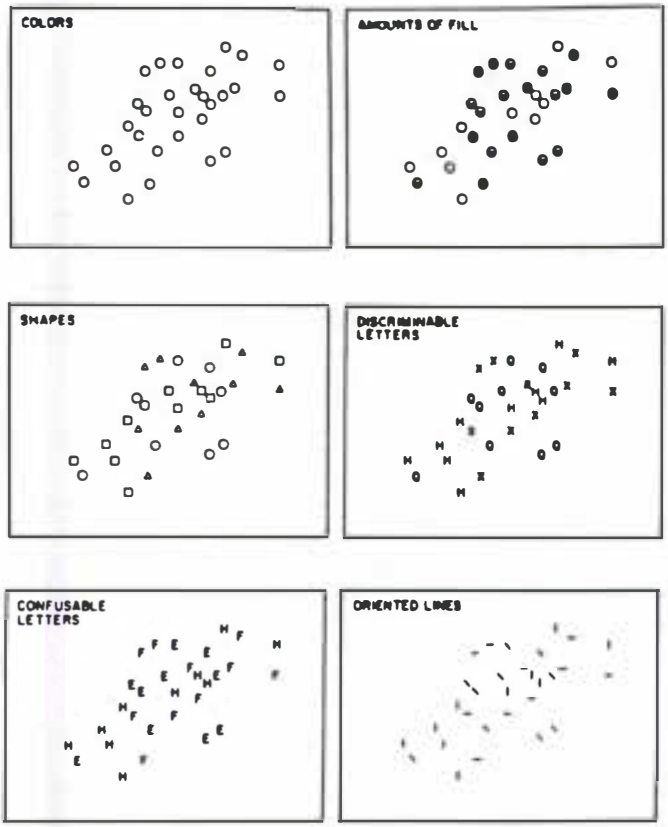


Fig.7 : Différents systèmes de discrimination de nuages de points (D'après Lewandowsky et Spence, 1989).

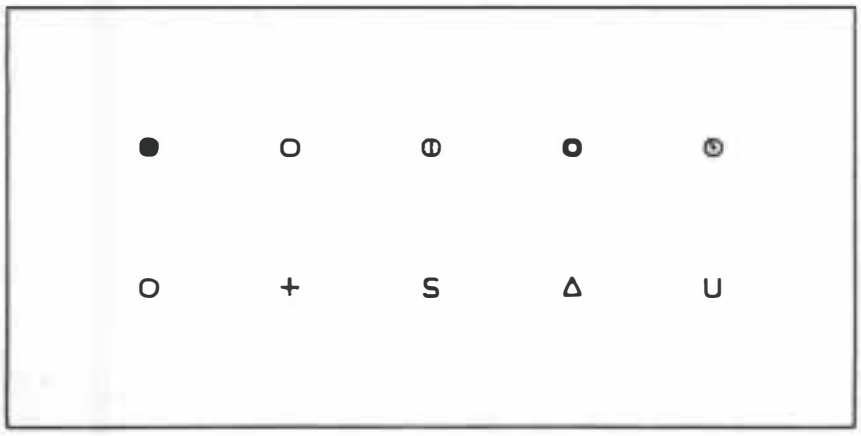


Fig.8 : Deux ensembles de motifs pour représenter les nuages de points. La première ligne de motifs est préférée quand il n'y a pas de problèmes de chevauchement de points (D'après Cleveland, 1985).

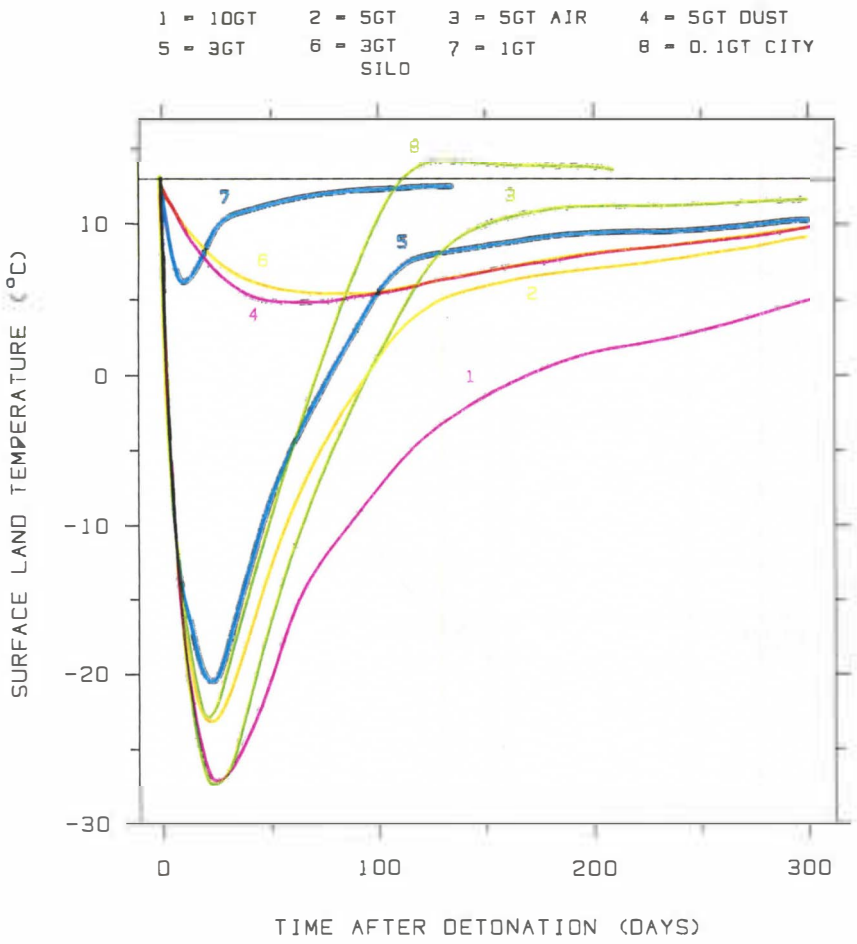


Fig.9 : Discrimination de courbes à l'aide de la couleur (D'après Cleveland, 1985).

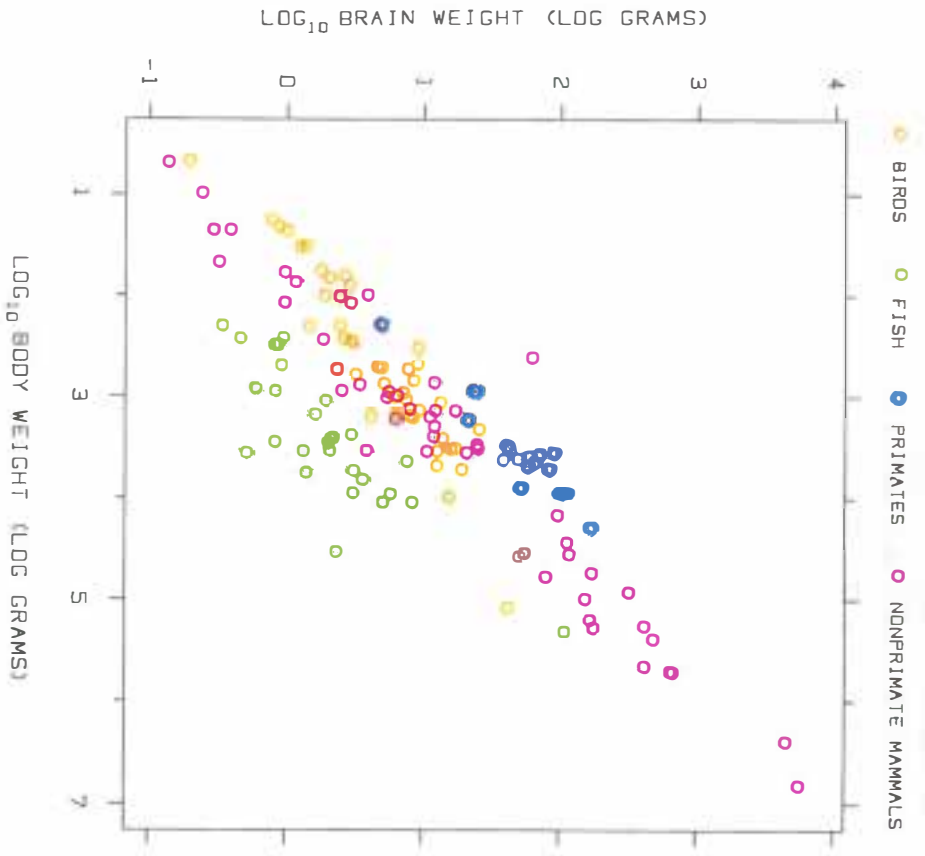


Fig.10 : Discrimination de nuages de points à l'aide de la couleur (D'après Cleveland, 1985)

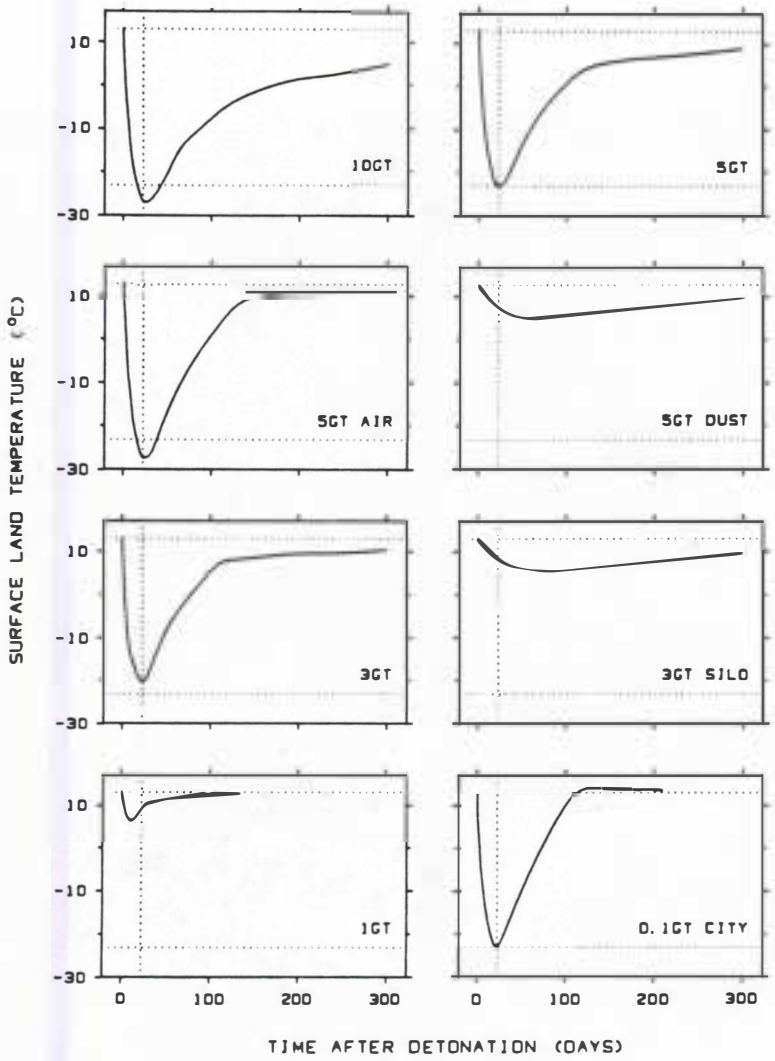


Fig.11 : Comparaison des courbes de la fig.5 par juxtaposition. Les droites horizontales des graphiques (marquées en pointillés) représentent la température moyenne de l'hémisphère nord et l'ordonnée du minimum de la courbe "5 GT" (D'après Cleveland, 1985).

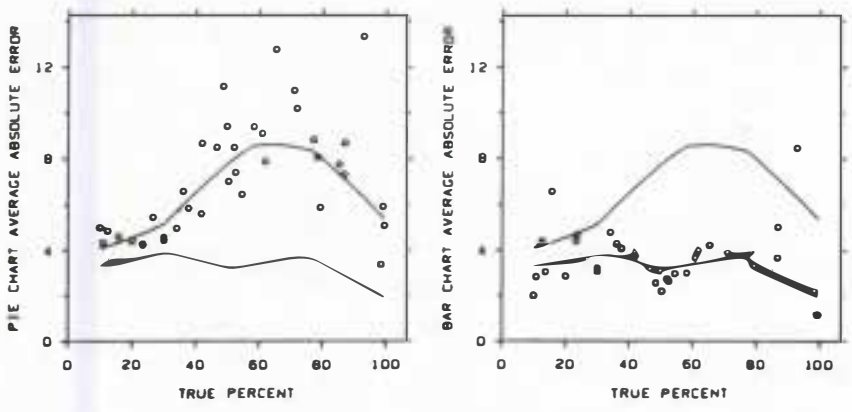


Fig.12 : Comparaison de deux nuages de points par juxtaposition. Les courbes représentées de manière identiques sur les deux graphiques correspondent aux lissages de chacun des nuages (D'après Cleveland, 1985).

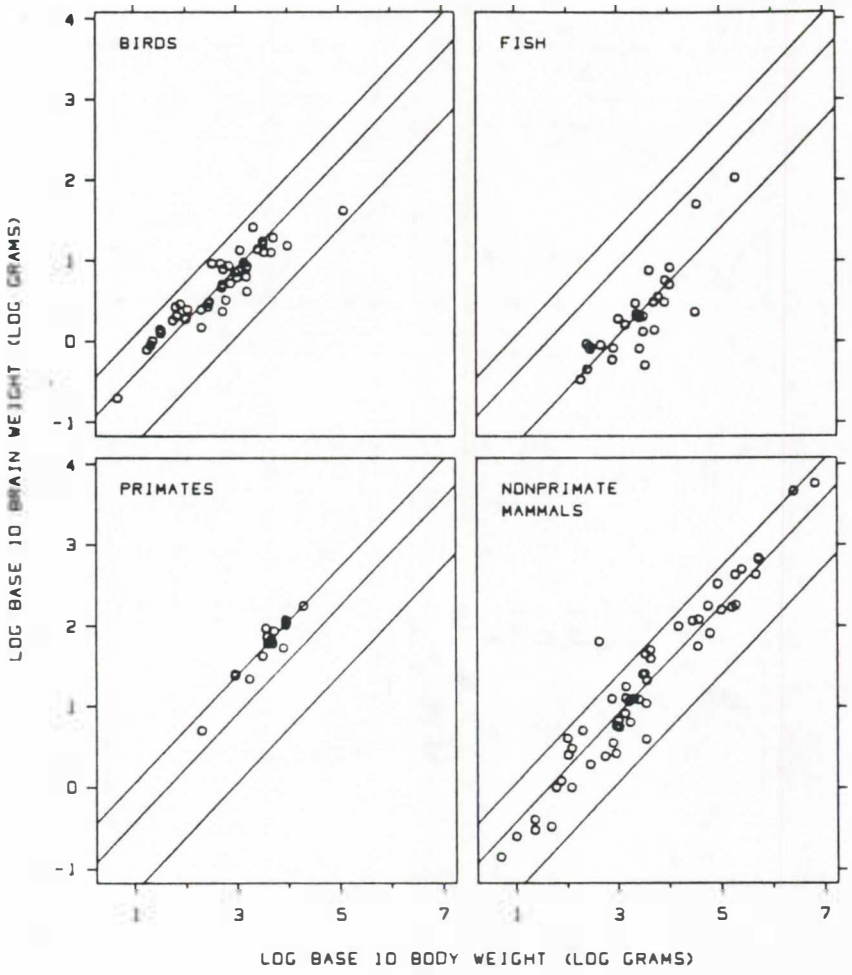


Fig. 13 : Comparaison de quatre nuages de points par juxtaposition. Les droites identiques sur chaque graphique correspondent aux régressions des nuages "primates", "birds", et "non primate mamals" (D'après Cleveland, 1985).

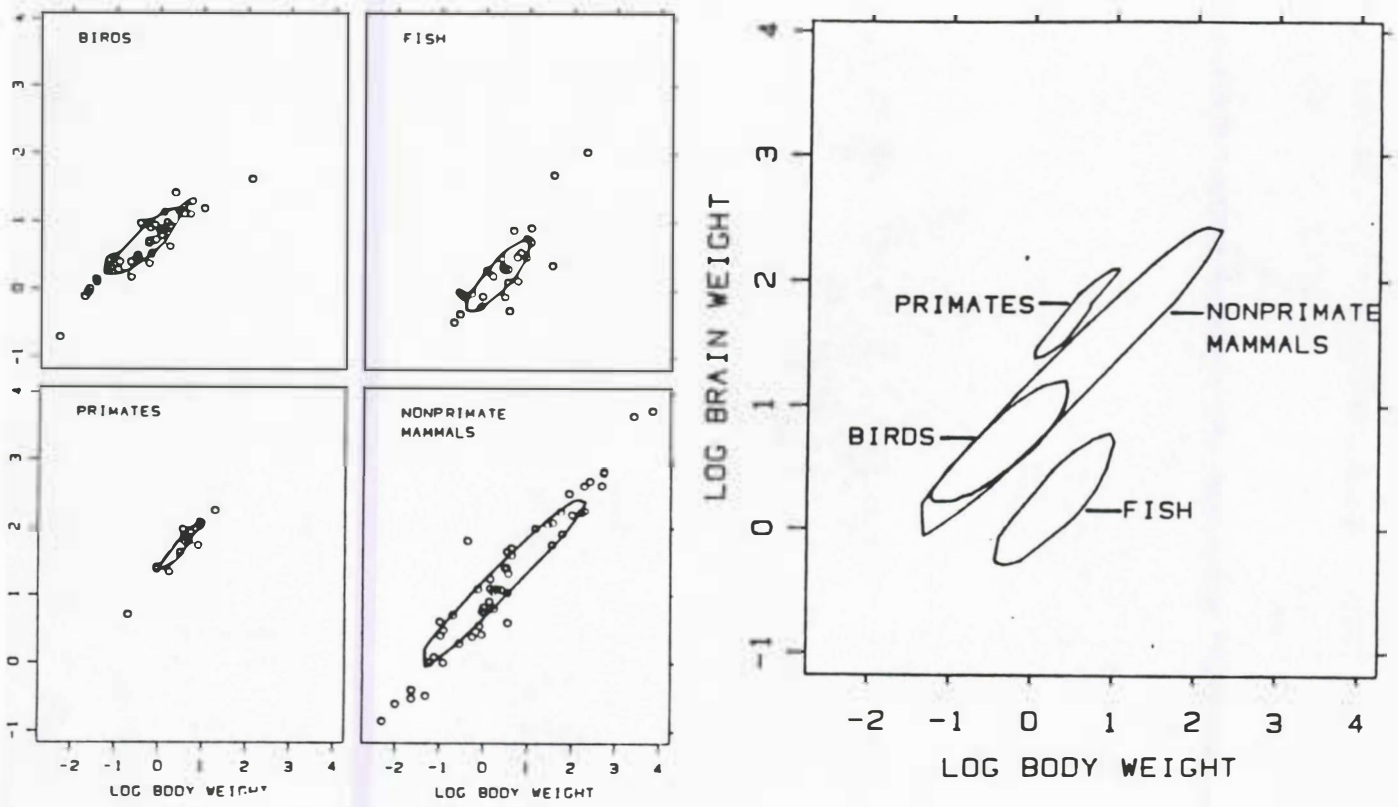


Fig. 14 : Comparaison de quatre nuages de points par lissages polaires (D'après Cleveland, 1984b).

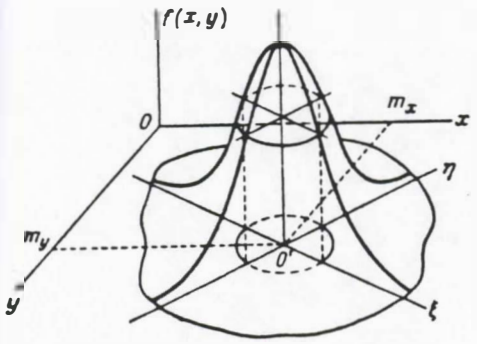


Fig. 15: Densité de probabilité d'une loi normale bidimensionnelle et ellipses de dispersion (D'après Ventsel, 1973).

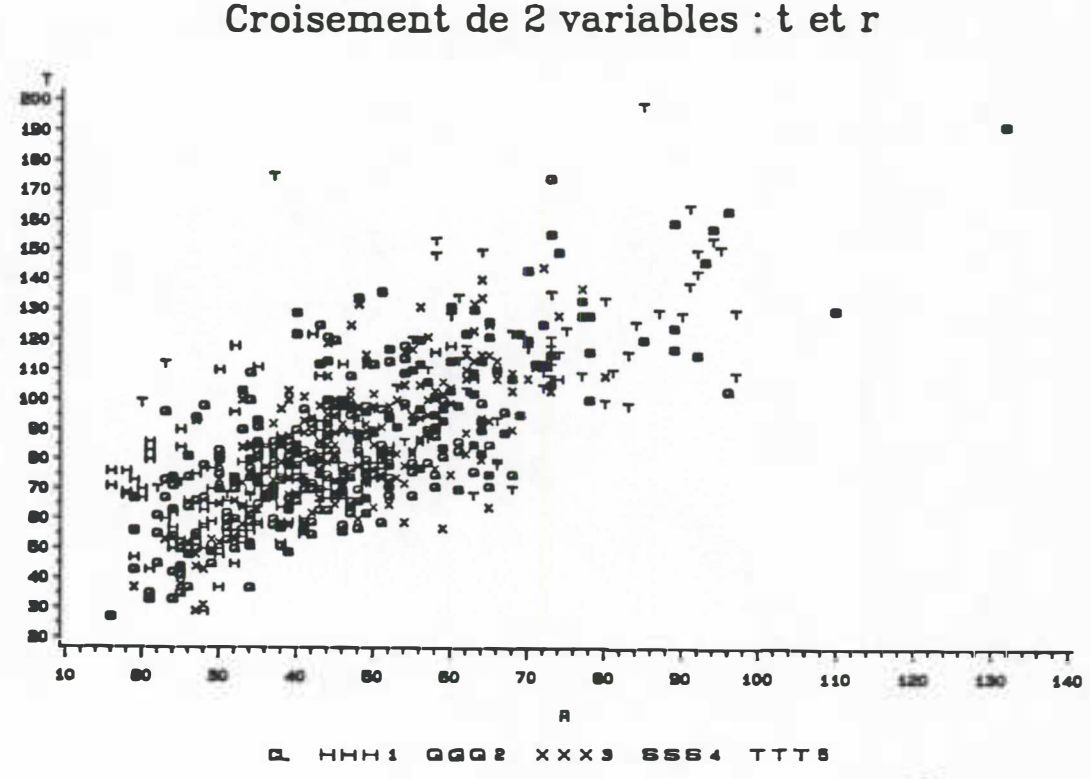


Fig. 16 : Retrait tangentiel (T) en fonction du retrait radial (R) pour cinq échantillons de bois de densités différentes. Chaque classe de densités est représentée par une lettre alphabétique différente. Le trop grand nombre de points rend le graphique illisible.

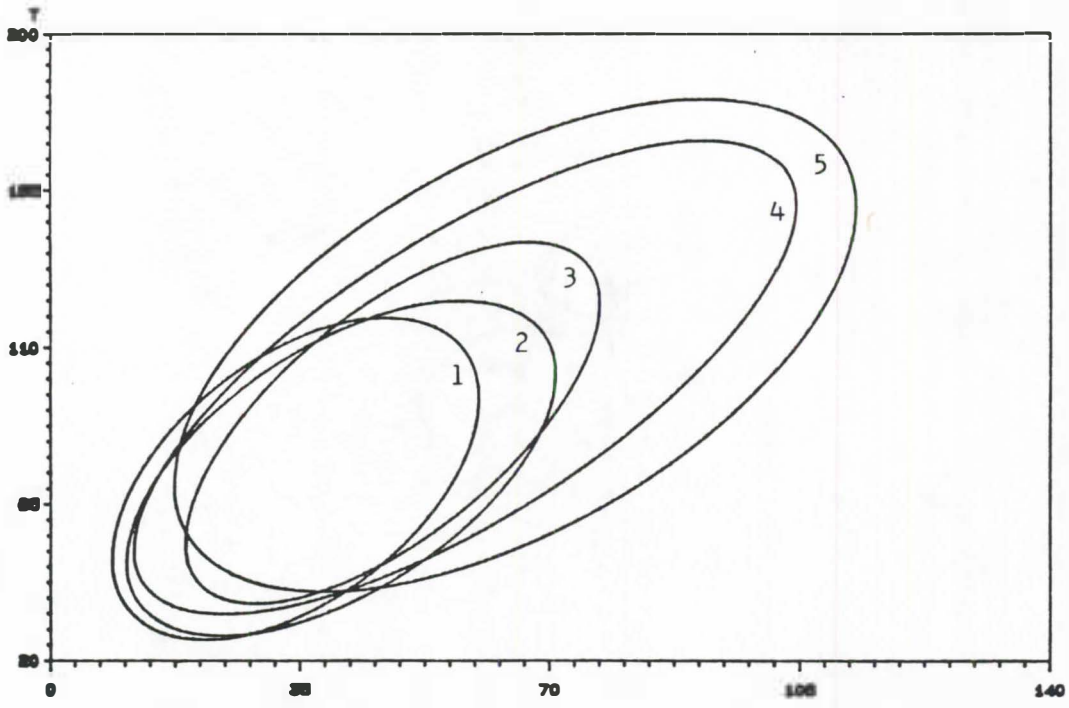


Fig. 17 : Ellipses de dispersion (p=0.95) des échantillons de bois de la figure n°16. Chaque ellipse contient environ 95% du nuage qu'elle représente.

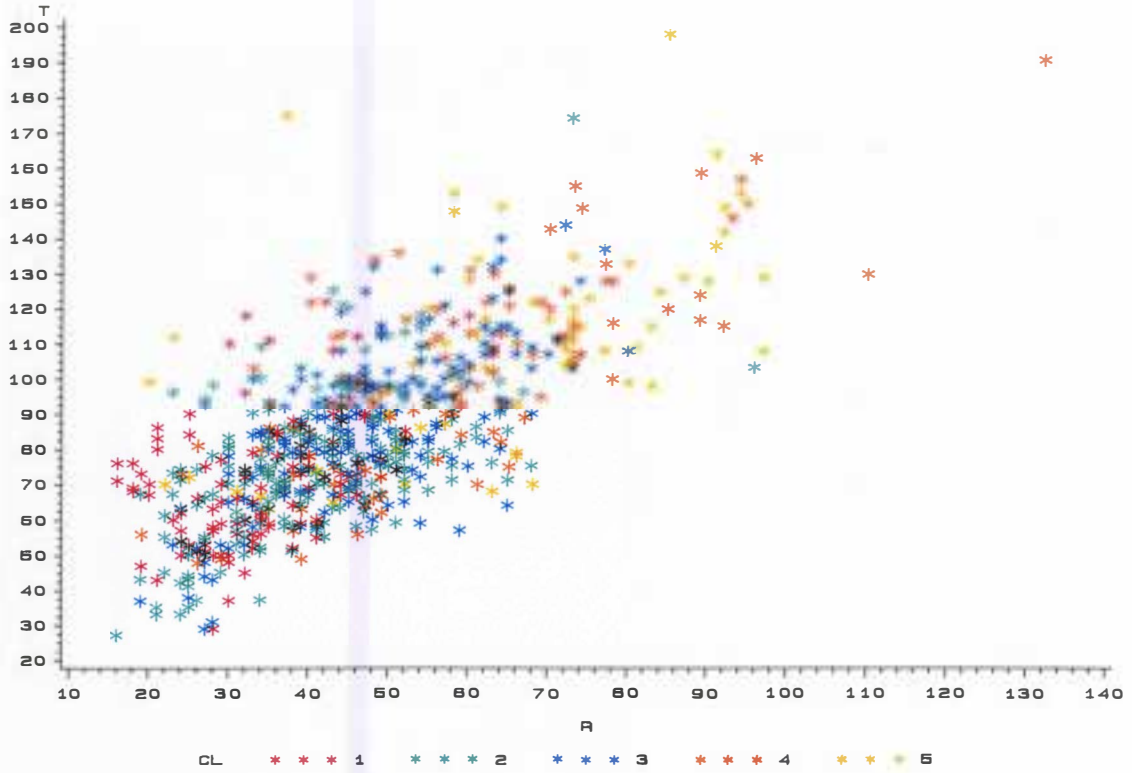


Fig. 18 : Retrait tangentiel (T) en fonction du retrait radial (R) pour cinq échantillons de bois de densités différentes. Chaque classe de densités est représentée par une couleur différente. Comme pour la figure n°16, le trop grand nombre de points rend le graphique peu lisible.

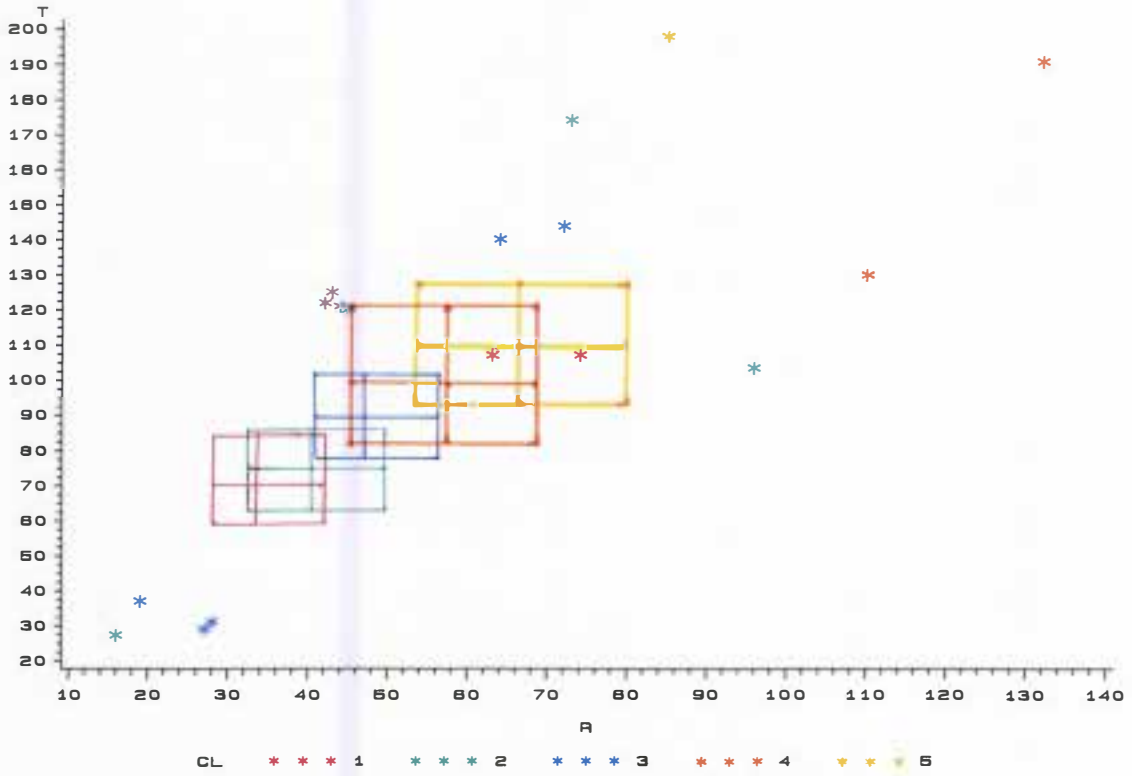


Fig.19 : *Box plots* bidimensionnels des échantillons de bois de la figure n°18. Les points extrêmes (*outliers*) ont été conservés pour chaque nuage.

CARTES INFORMEES

Septembre 1992

Jean-Claude BERGONZINI

BIOMETRIE
CIRAD - Forêt

BIOM/JCB/AG

CARTES INFORMEES

1. DES OUTILS DE REPRESENTATION

La carte présentée ci-dessous *(Fig. 1)* illustre les taux de mortalité par meurtre aux Etats-Unis. Ce taux est égal au rapport du nombre de meurtres commis en 1978 sur le nombre d'habitants de l'état concerné. Le taux est donné pour 100.000 habitants.

Ce type de représentation est très fréquemment utilisé lorsqu'on est en présence d'une variable quantitative évoluant dans un espace géographique discrétisé.

Les deux défauts essentiels de cette cartographie sont les suivants :

a) L'impact visuel est associé à la superficie des états, laquelle n'est pas nécessairement corrélée avec le nombre d'habitants ;

b) la symbolisation utilisée induit une partie d'informations non négligeable et l'utilisation de la trame peut conduire à des effets de grisé non désirables.

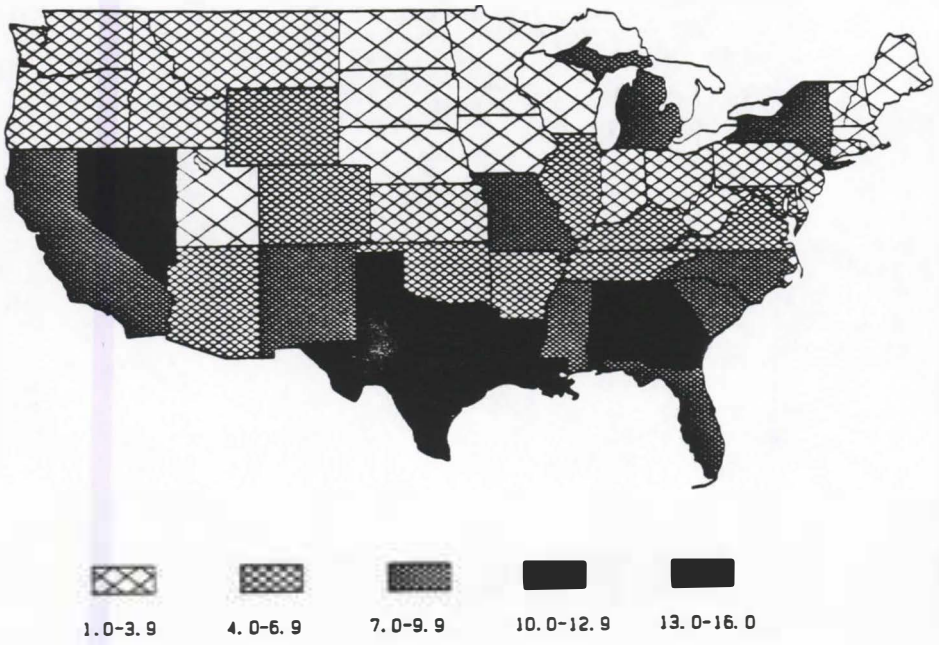


Fig. 1

.../...

Une première alternative a été proposée par CLEVELAND et McGILL qui se font les avocats de l'utilisation d'"indicateurs de niveau" *(Fig. 2 et 3)*.

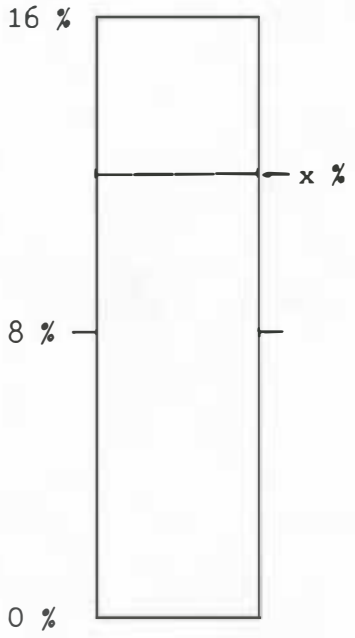


Fig. 2 - *L'indicateur de niveau*
(donné en parts de 10.000)

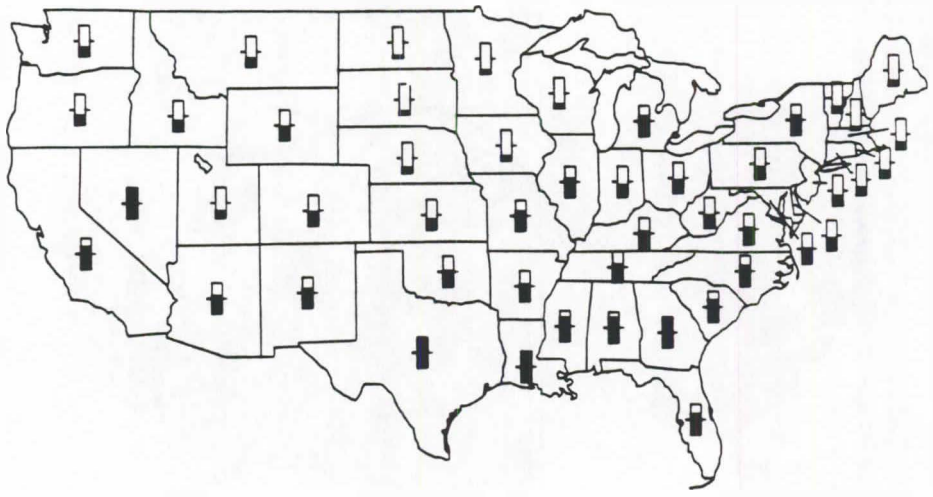


Fig. 3

La hauteur du rectangle est proportionnelle au taux de meurtres.

Exemple :

.../...

L'inconvénient de cette nouvelle approche est liée au type d'information mobilisé (un pourcentage) dont on peut juger des limites en comparant les cartes des *Fig. 3 et 4*.

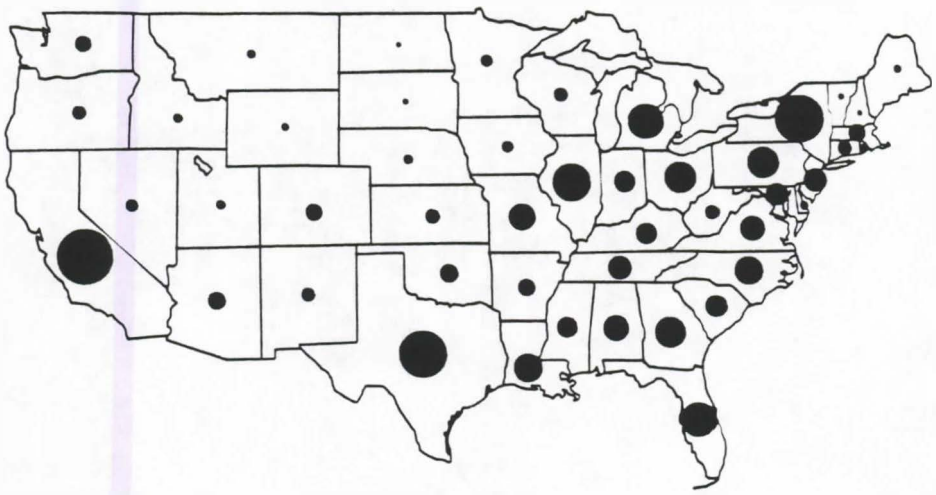


Fig. 4

La grandeur du cercle est proportionnelle au nombre de meurtres.

Exemple :

.../...

R. DUNN propose d'utiliser la largeur des niveaux utilisés par CLEVELAND comme indicateur de l'importance des populations concernées.

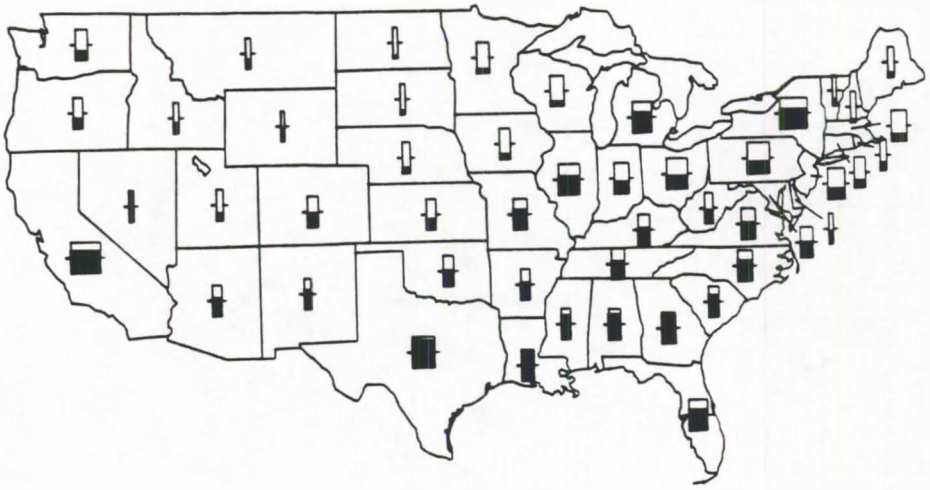


Fig. 5

La hauteur du rectangle est proportionnelle au taux de meurtres ; la largeur est proportionnelle à la racine carrée de la densité de population.

Exemple :

Taux de meurtres
Population

.../...

2. COMPARAISONS ENTRE DEUX REPRESENTATIONS

Les représentations concernées sont celles des *Fig. 1 et 3*, donc des *cartes C1 et C2*. Les données sont échelonnées pour varier de 0 à 48 et non plus de 0 à 16. (En fait, 48 est aussi égal au nombre d'états et le taux de meurtres τ suit approximativement une loi uniforme sur l'intervalle 0-48).

2.1. Le Protocole

a) 50 juges sont utilisés. Ils reçoivent une courte formation.

b) Les 48 états sont répartis en deux groupes G_1 et G_2 comparables (les critères de comparaison sont la superficie et τ).

c) On peut procéder selon quatre itinéraires :

Itinéraire	on observe → on évalue → on observe → on évalue
I_1	C_1 → G_1 → C_2 → G_2
I_2	C_1 → G_2 → C_2 → G_1
I_3	C_2 → G_1 → C_1 → G_2
I_4	C_2 → G_2 → C_1 → G_1

d) Les juges sont répartis en quatre sous-groupes J_1, J_2, J_3, J_4. Chaque groupe effectue un itinéraire (répartition au hasard). Chaque juge doit évaluer τ pour chaque état. Il y a donc au total 2.400 estimations.

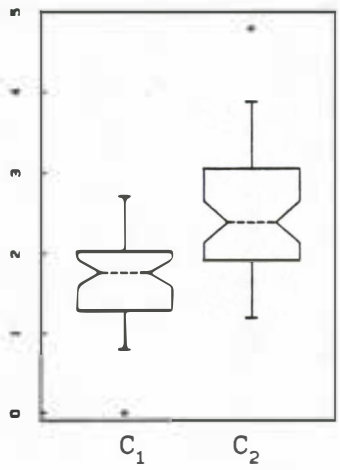
2.2. Les résultats

Pour chaque juge et chaque estimation (chaque état), on calcule l'écart entre la valeur estimée et la vraie valeur. On dispose ainsi, pour chaque juge, de deux séries de séries :

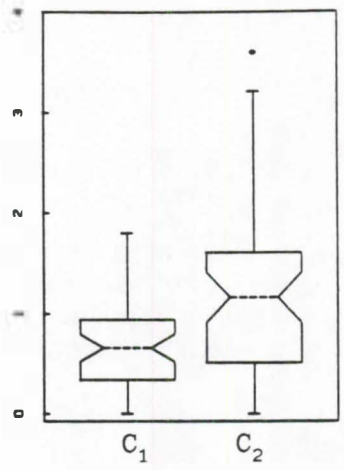
Juge J —
- x_1 , x_2 , ... x_{24} *(Carte C_1)*
- x'_1, x'_2, ... x'_{24} *(Carte C_2)*

Les résultats sont synthétisés sur les graphiques des *Fig. 6 et 7*.

.../...

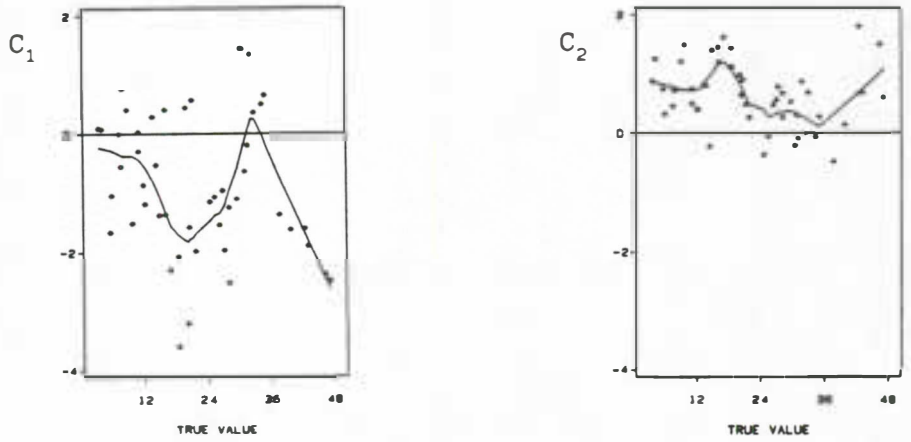


Ecarts types

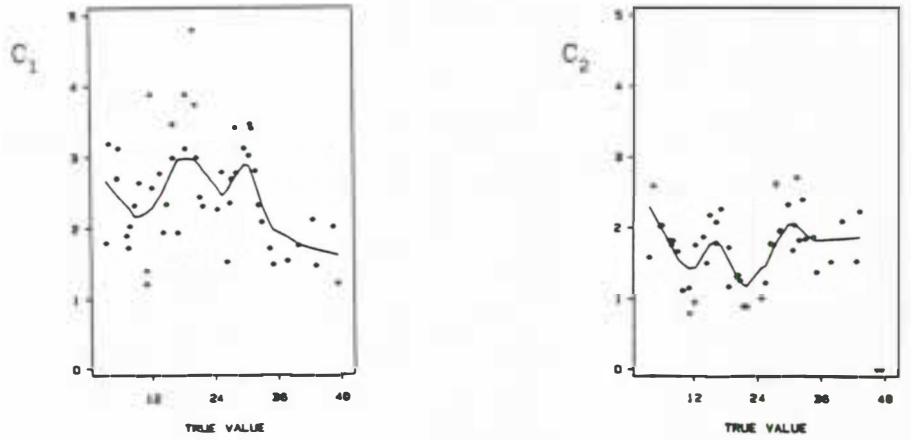


Moyenne des erreurs absolues

Fig. 6 : Boîtes à moustaches avec entailles



Moyenne des erreurs absolues



Ecarts types

Fig. 7

.../...

Bibliographie

. The American Statistician - May 1987 - Vol. 41, n° 2

. The American Statistician - May 1988 - Vol. 42, n° 2.

* * *

FAUT-IL SE MEFIER DES GRAPHIQUES ?

Septembre 1992

Matthieu LESNOFF

BIOMETRIE
CIRAD - Forêt

FAUT-IL SE MEFIER DES GRAPHIQUES ?

1. Les principes de représentation graphique de Tufte.

Dans son ouvrage sur les méthodes de représentation graphique, Tufte (1983) présente une réflexion sur certains pièges graphiques pouvant tromper facilement le lecteur non averti. Selon lui, un bon graphique ne doit pas créer de distorsion entre la représentation visuelle des données et les quantités numériques que celles-ci représentent. Mais l'auteur pose le problème de fond du sens réel de l'expression "représentation visuelle" : Est-ce la mesure physique d'un élément graphique (longueur, surface, etc.) ? Ou l'effet visuel perçu ? Comment savoir si l'image représente fidèlement les données sous-jacentes ?

De nombreuses expériences sur la perception visuelle des formes et des figures ont été mises en place pour tenter de répondre à ces questions, et constituent une littérature abondante. Elles se basent toutes sur le même principe : des lignes de longueur variable, des cercles de surface variable, etc. sont présentés à des sujets dont les appréciations sur ces variations sont enregistrées sous forme numérique.

Malheureusement, ces expériences n'ont pas abouti à des lois générales quantitatives et les résultats ont été assez décevants. En fait, les chercheurs se sont aperçus que la perception d'un élément graphique et de ses variations dépendait fortement du contexte de l'expérience (sujet interrogé, type de forme de l'élément graphique, etc.). Faut-il alors construire un graphique pour chaque contexte ? Ou construire un graphique pour un contexte "moyen" ?

Devant ces difficultés, et dans l'espoir de représenter les données le plus fidèlement possible, Tufte propose de se limiter à essayer d'acquérir une certaine uniformité dans les représentations graphiques, en respectant plusieurs principes d'intégrité :

- Si les données quantitatives sont représentées par des mesures de longueurs, de surfaces ou de volumes d'éléments graphiques, celles-ci doivent être directement proportionnelles aux quantités numériques représentées.

- Une légende claire, détaillée et approfondie doit être constituée afin d'éliminer les ambiguïtés et les problèmes d'interprétation éventuels dus à des distorsions ou à d'autres effets graphiques. Les évènements importants contenus dans les données doivent être inscrits dans la légende.

- Une représentation graphique doit mettre en valeur les variations des données, et non des variations artificielles dues à des changements d'échelle.

- Le nombre de dimensions graphiques (variables) ne doit pas dépasser le nombre de dimensions contenues dans les données.

- Le graphique ne doit pas extraire les données de leur contexte.

Tufte définit un facteur d'erreur relativement au premier des principes présentés, dont le non respect est une cause très fréquente de mauvaise interprétation graphique :

$$F = \frac{\text{mesure de l'effet au niveau du graphique}}{\text{mesure de l'effet réel (données)}} \times 100$$

Si F=1, le graphique a globalement de fortes chances de retranscrire fidèlement les données numériques. Si F<0.95 ou F>1.05, la distorsion graphique introduite est substantielle.

2. Quelques exemples de mauvaises représentations graphiques.

Tufte (1983) et Wainer (1984) donnent un ensemble d'exemples de mauvaises représentations graphiques. Ils sont repris ici de manière quasi intégrale. Un premier exemple permet d'illustrer le calcul du facteur d'erreur de Tufte.

exemple n°1 (Tufte, 1983) :

La figure n°1a présente l'évolution de la consommation moyenne en carburant (miles/gallon) des automobilistes américains. Pour chaque année, la distance moyenne parcourue par unité de carburant est représentée par une ligne horizontale perpendiculaire à l'axe de la route.

De 1978 à 1985, la distance parcourue par les automobilistes a augmenté de : 100(27.5-18.0)/18.0=53%. Sur le graphique, cette variation de distance est perçue à l'aide des tailles relatives des deux lignes correspondantes, soit d'après le dessin original : 100.(5.3-0.6)/0.6=783%. Le facteur d'erreur est donc : F=783/53=14.8.

En plus de cette distorsion importante, Tufte souligne d'autres effets de perspective néfaste à la perception globale du graphique :

- le futur est souvent représenté en arrière plan, vers l'horizon ; le graphique inverse cette convention ;
- les chiffres identifiant les dates à gauche de la route restent de même taille, alors que ceux identifiant les distances ont des tailles de plus en plus grandes ;
- de même que la longueur des lignes, l'évolution de la taille des chiffres de droite n'est pas proportionnelle à l'évolution réelle des données.

Tufte propose un graphique plus simple que le précédent graphique "élaboré", mais qui permet d'apprécier beaucoup mieux l'évolution réelle des données numériques (fig.1b).

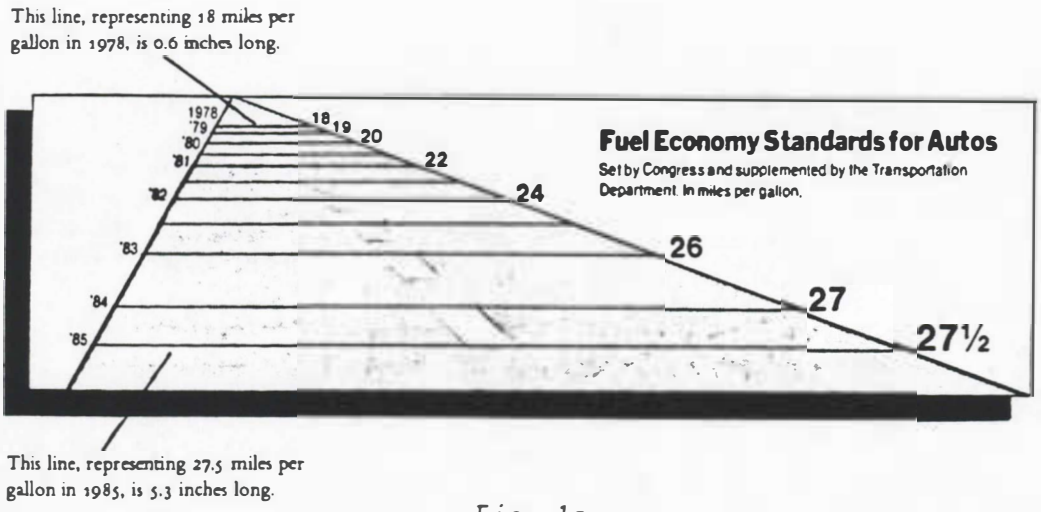


Fig. 1a

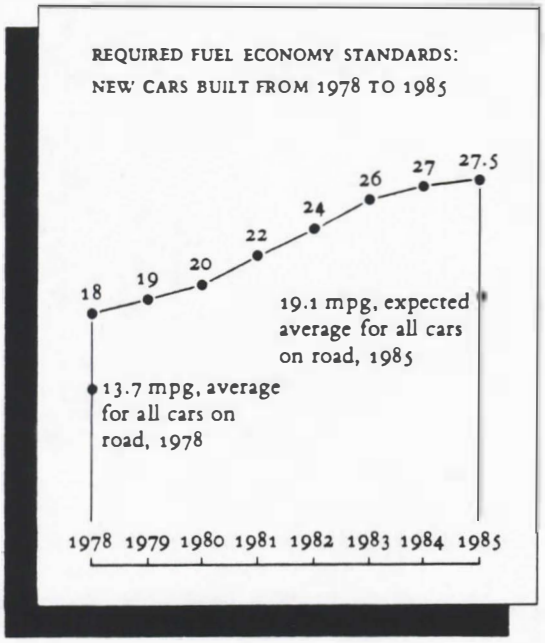


Fig. 1b

Les trois exemples suivants concernent le problème de la présence de plusieurs échelles différentes dans un même graphique. L'utilisation d'échelles multiples est à éviter (3^{e} principe) car elle entraîne presque toujours une mauvaise appréciation des données sous-jacentes. En général, le lecteur éprouve en effet beaucoup de difficultés à percevoir les changements d'échelles dans un même graphique, et à en apprécier les conséquences au niveau des données : "l'oeil attend une échelle régulière du début jusqu'à la fin du graphique" (Tufte, 1983).

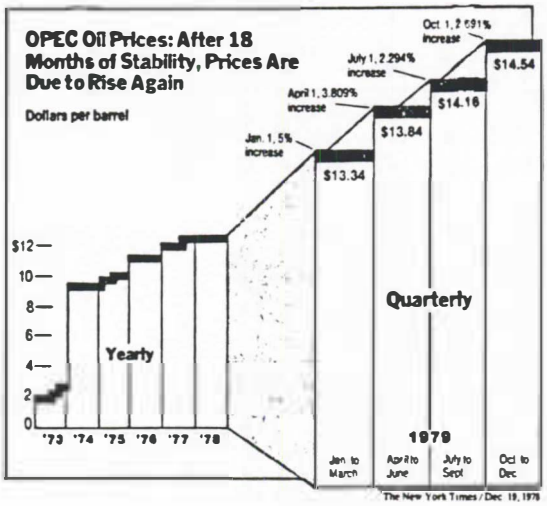
exemple n°2 (Tufte, 1983) :

Le graphique de la figure n°2 présente l'évolution du prix de l'huile de 1973 à 1979. Il se compose de deux parties : la partie gauche est graduée en années (1973 à 1978) ; la partie droite en périodes de trois mois (1979). De même que l'échelle horizontale, l'échelle verticale est variable, ce qui multiplie les effets de distorsion entre les deux parties du graphique :

dans la partie gauche : 10$=0.31 inches^2
dans la partie droite : 10$=4.69 inches^2

La même quantité (10$) est donc représentée de manière 4.69/0.31=15.1 fois plus forte à droite qu'à gauche. Il est ainsi très difficile de mettre en relation les deux parties du graphique et d'apprécier l'évolution réelle des données jusqu'en 1979. La pente observée lors de l'augmentation du prix de l'huile en 1979 n'a absolument rien à voir avec les données numériques originales.

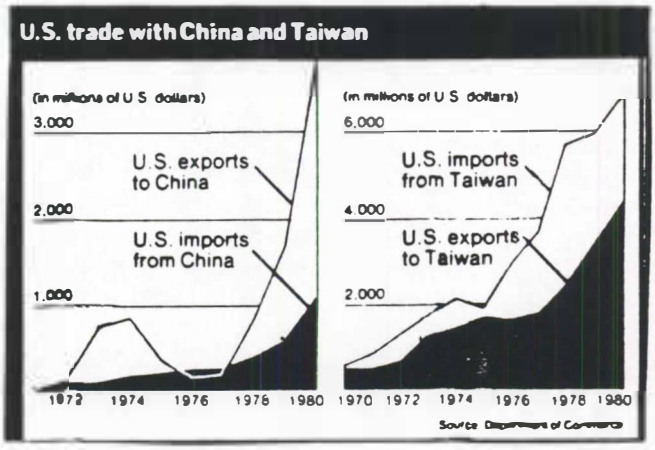
Fig. 2



exemple n°3 (Wainer, 1984) :

Le graphique de la figure n°3 présente les mêmes difficultés d'interprétation que celui de la figure n°2 : il juxtapose deux échelles verticales différentes (1 à 3000 et 1 à 6000) et fausse la comparaison des variables quantitatives étudiées. Son autre défaut est que la surface ombrée représente les importations dans sa partie gauche et les exportations dans sa partie droite.

Fig. 3



exemple n°4 (Wainer, 1984) :

La figure n°4a donne un dernier exemple du danger des changements d'échelles. L'évolution du revenu des physiciens paraît linéaire, mais cette impression n'est due qu'à un effet d'échelle : celle-ci débute en effet par des intervalles de 8 ans et s'achève par des intervalles annuels. L'utilisation d'une échelle de temps régulière montre une tout autre représentation de l'évolution des revenus (fig.4b).

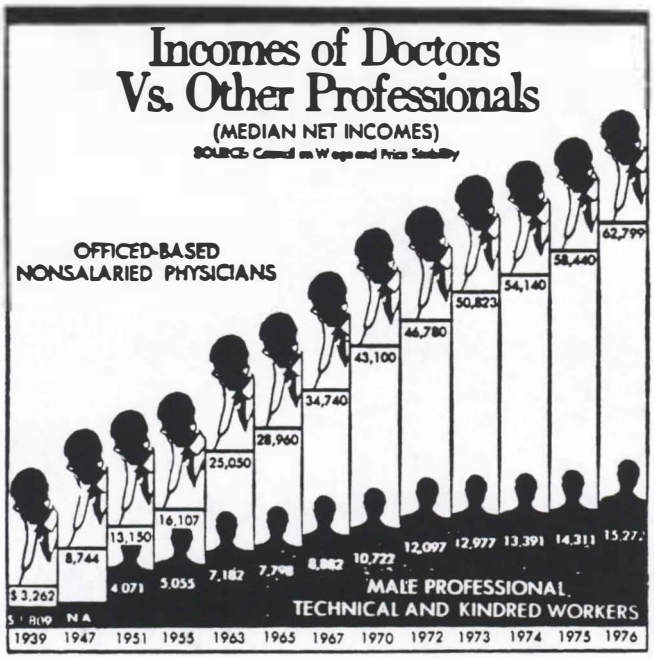


Fig. 4a

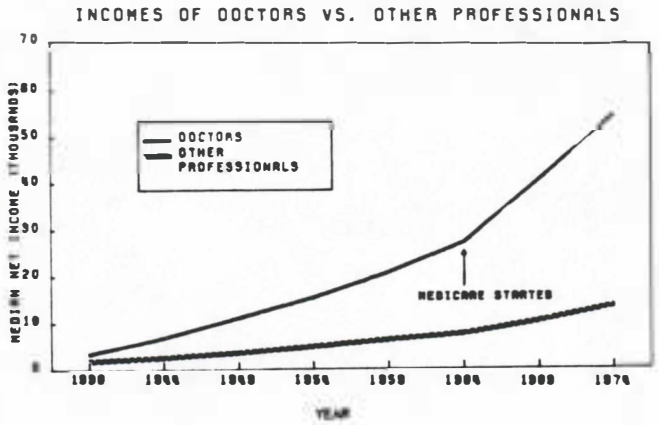


Fig. 4b

Tufte et Wainer ont souligné l'inefficacité et la faiblesse des représentations de variables unidimensionnelles par des surfaces ou des volumes : en général, les variations de surfaces ou de volumes ne sont pas proportionnelles aux variations réelles des données. Plusieurs exemples montrent que l'introduction de dimensions superflues dans un graphique peut, d'une part, entraîner des distorsions importantes, et d'autre part, diminuer la lisibilité du graphique.

exemple n°5 (Tufte, 1983 et Wainer, 1984) :

La figure n°5a décrit le taux d'inflation américain entre 1958 et 1978. La représentation du prix du dollar par des surfaces crée une distorsion d'environ 300% sur toute la période d'étude (F≈3). Wainer propose un graphique plus simple et beaucoup plus fidèle à l'évolution réelle du dollar (fig.5b).

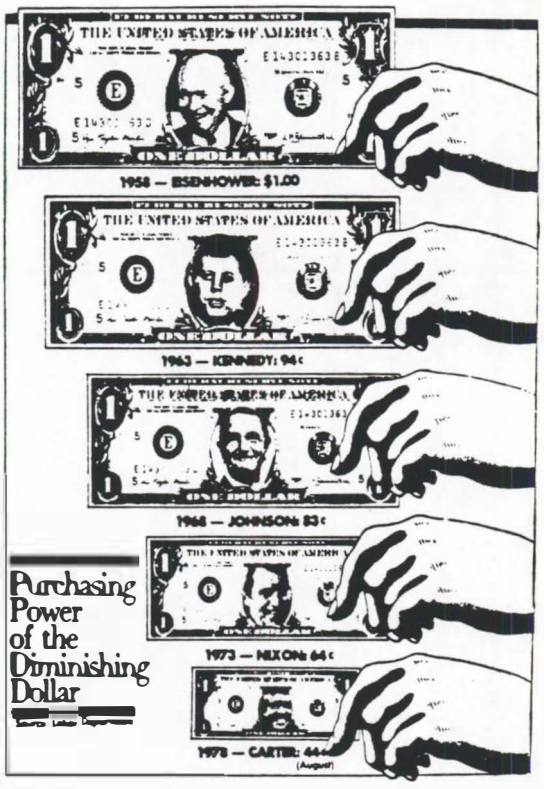


Fig. 5a

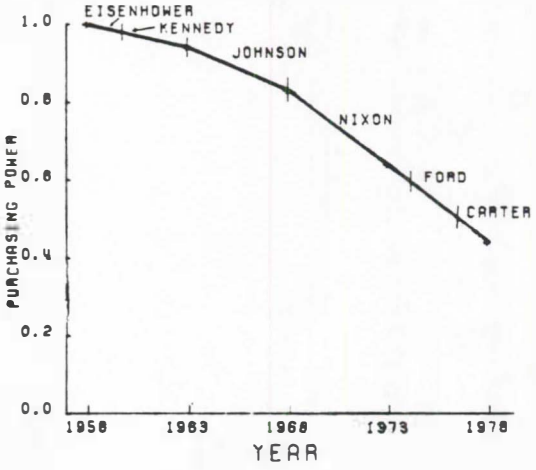


Fig. 5b

exemple n°6 (Tufte, 1983) :

Les "cercles de W. Playfair" (fig.6), qui représentent la taille des populations urbaines (données unidimensionnelles) par des surfaces, ont le même défaut que le graphique précédent.

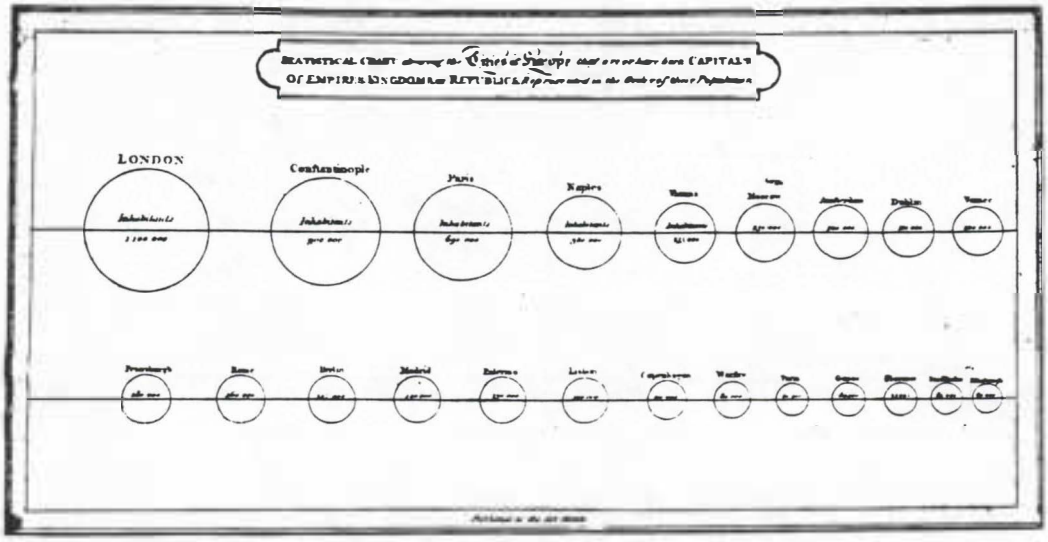
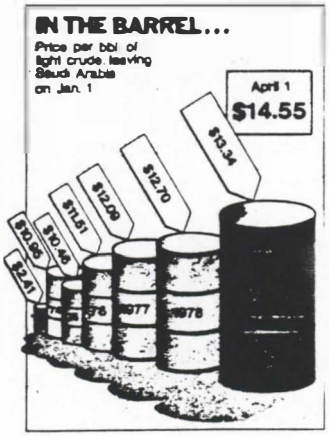


Fig. 6

exemple n°7 (Tufte, 1983) :

Le problème de non proportionnalité est encore amplifié lorsqu'on utilise non plus des surfaces mais des volumes. Pour le graphique de la figure n°7, le facteur d'erreur est F=9.4 si l'on considère la surface des barils, et F=59.4 si l'on considère leur volume. Il faut aussi remarquer que ce graphique a une échelle horizontale variable (comme la plupart des graphiques utilisant des représentations surfaciques ou volumiques) puisque les barils ne sont pas espacés de manière régulière.

Fig. 7



exemple n°8 (Wainer, 1984) :

Wainer fait un parallèle intéressant entre un tableau de données dont les chiffres contiennent un trop grand nombre de décimales, qui nuisent à la lisibilité de l'ensemble, et un graphique contenant des dimensions superflues. L'information contenue dans les colonnes du tableau n°1a (nombres à 5 décimales) est très peu accessible au lecteur. Dans le tableau n°1b, on voit par exemple sans difficulté que la 3^{e} et la 5^{e} colonnes sont très semblables à un facteur 10 près.

Tab. 1a

	b/c = 10.0		100.0		1,000.0	
N	r^*	$(G_M(r^*) - a)/c$	r^*	$(G_M(r^*) - a)/c$	r^*	$(G_M(r^*) - a)/c$
3	2	.20000	2	2.22500	2	22.47499
4	2	.26333	2	2.88833	2	29.13832
5	2	.32333	3	3.54167	3	35.79166
6	3	.38267	3	4.23767	3	42.78764
7	3	.44600	3	4.90100	3	49.45097
8	3	.50743	4	5.57650	4	56.33005
9	3	.56743	4	6.26025	4	63.20129
10	4	.62948	4	6.92358	4	69.86462

NOTE: $g(Xs + r - 1) = bR(Xs + r - 1) + a$, if $S = s$, and $g(Xs + r - 1) = 0$, otherwise.

Tab. 1b

	b/c = 10		b/c = 100		b/c = 1,000	
N	r^*	G	r^*	G	r^*	G
3	2	.2	2	2.2	2	22
4	2	.3	2	2.9	2	29
5	2	.3	3	3.5	3	36
6	3	.4	3	4.2	3	43
7	3	.4	3	4.9	3	49
8	3	.5	4	5.6	4	56
9	3	.6	4	6.3	4	63
10	4	.6	4	6.9	4	70

NOTE: $g(Xs + r - 1) = bR(Xs + r - 1) + a$, if $S = s$, and $g(Xs + r - 1) = 0$, otherwise.

De la même manière qu'un nombre important de décimales peut rendre un tableau difficile à comprendre, l'augmentation du nombre de dimensions d'un graphique peut le rendre complètement confus (attention aux histogrammes univariés en 3-D...). Il semble même, sur le graphique de la figure n°8a, que cela ait trompé le dessinateur lui-même : l'année "1975" du graphique correspond en fait à la troisième dimension de l'année 1976. Le dessinateur a rajouté un intervalle fictif et il faudrait décaler les années 1975, 1977 et 1978 pour retrouver les valeurs correctes. L'évolution des deux variables étudiées est bien plus lisible sur le graphique proposé par Wainer (fig.8b).

Fig. 8a

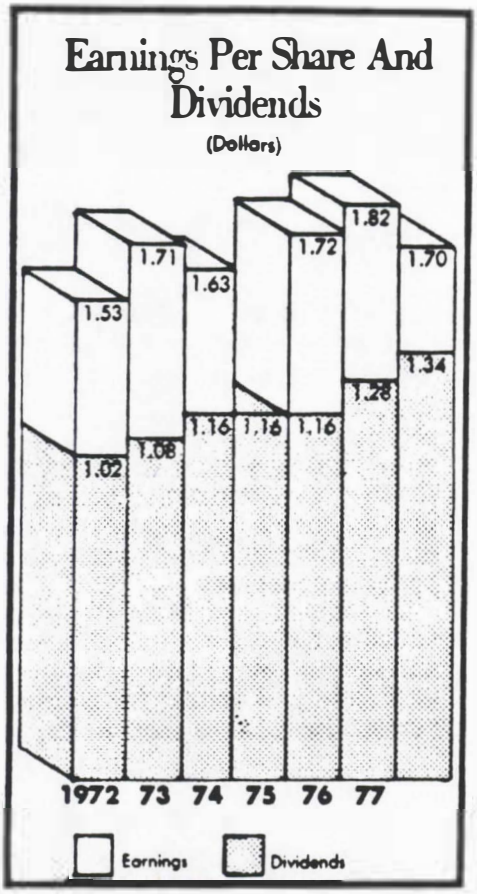
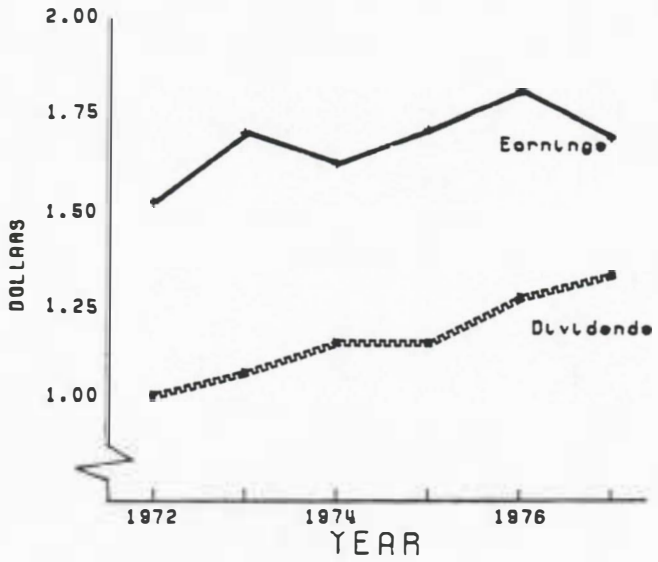

Earnings Per Share And Dividends
(Dollars)
1.53
1.71
1.63
1.72
1.82
1.70
1.02
1.08
1.16
1.16
1.16
1.28
1.34
1972
73
74
75
76
77
Earnings
Dividends

Fig. 8b


2.00
1.75
1.50
1.25
1.00
DOLLARS
Earnings
Dividends
1972
1974
1976
YEAR

Selon Wainer, un bon graphique ne doit pas cacher l'information contenue dans les données et doit la faire ressortir de manière pertinente. Il présente plusieurs graphiques qui justement ne mettent pas en valeur cette information (sûrement involontairement) du fait d'une mauvaise organisation graphique.

exemple n°9 (Wainer, 1984) :

Le graphique de la figure n°9a a le défaut de masquer la croissance des écoles privées de 1929 à 1970, en utilisant une échelle trop grande par rapport à cette variable. Une échelle plus réduite permet de mieux apprécier l'augmentation du nombre d'écoles privées.

Fig. 9a

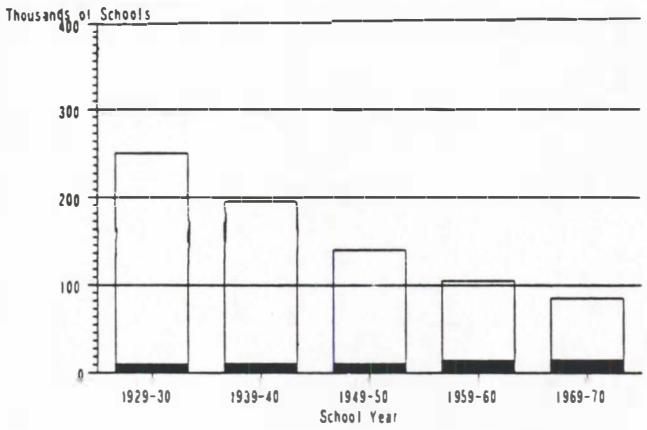
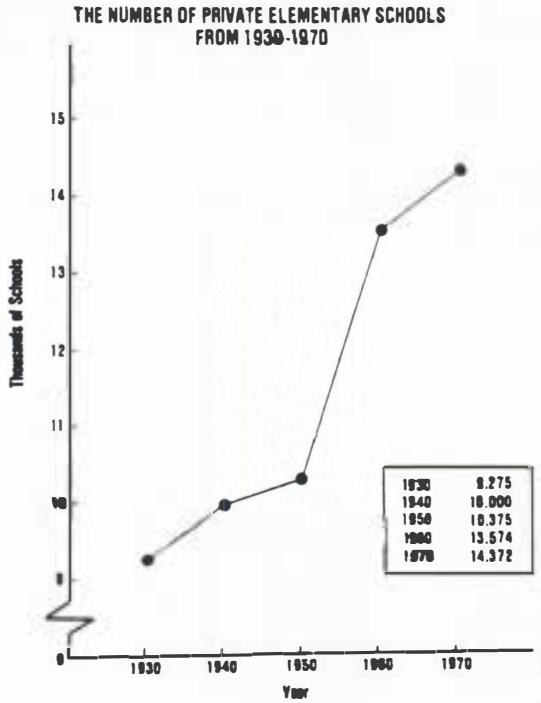


Fig. 9b



exemple n°10 (Wainer, 1984):

Il est fréquent que les données étudiées contiennent certaines informations triviales et d'autres plus intéressantes. Un graphique peut devenir de mauvaise qualité s'il ne met en valeur que l'aspect trivial des données. Ainsi, le graphique de la figure n°10a souligne le fait très connu de l'augmentation du salaire avec le niveau d'étude. Il cache cependant le fait plus intéressant d'une différence de salaires entre hommes et femmes à niveau d'étude égal. Cet aspect apparaît clairement dans le graphique de la figure n°10b.

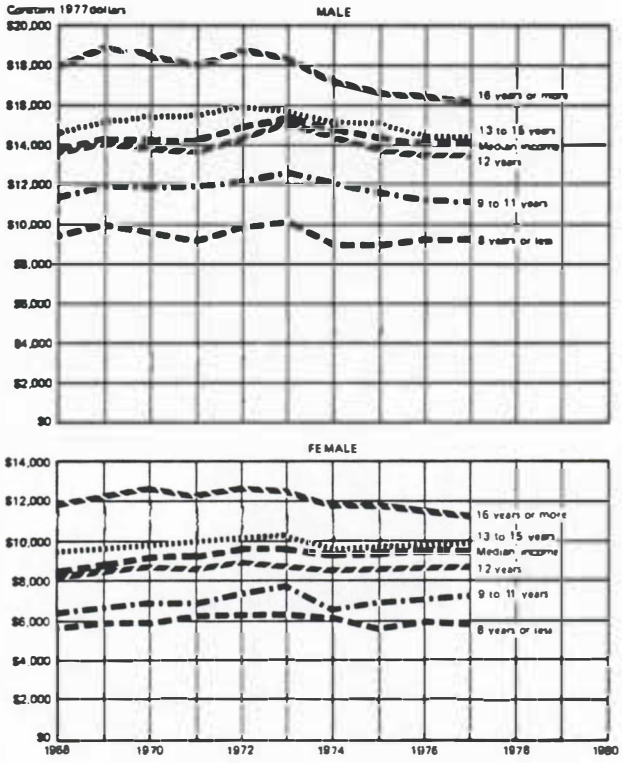


Fig. 10a

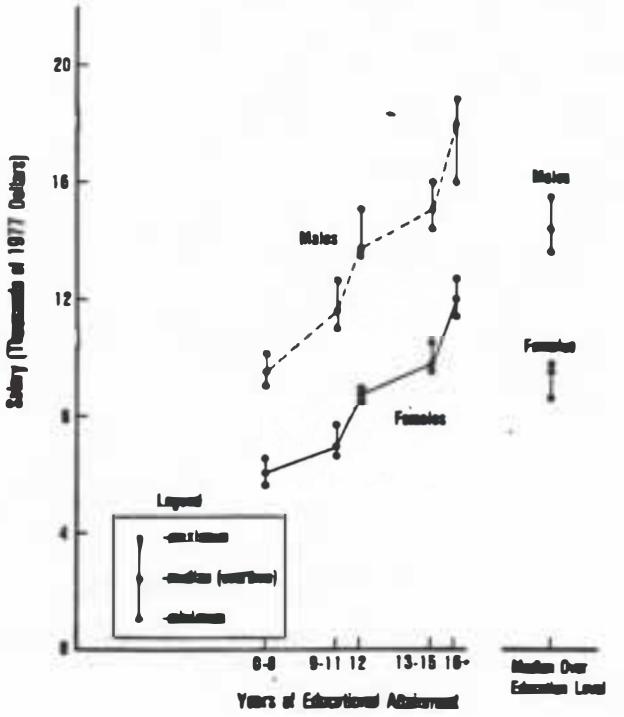


Fig. 10b

exemple n°11 (Wainer, 1984) :

Le graphique de la figure n°11a, qui représente les espérances de vie par sexe de différents pays ordonnés par ordre alphabétique, apparaît bien confus et beaucoup moins facilement interprétable que le graphique *stem and leaf* de la figure n°11b. Ce dernier graphique met tout de suite en évidence la différence d'espérance de vie entre les hommes et les femmes, le classement des pays ainsi que ceux qui se détachent le plus (ex : URSS).

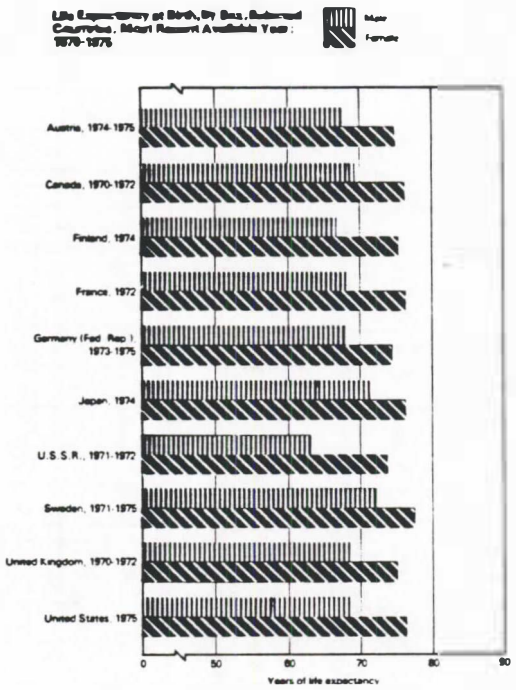


Fig. 11a

LIFE EXPECTANCY AT BIRTH, BY SEX,
MOST RECENT AVAILABLE YEAR

WOMEN	YEARS	MEN
SWEDEN	78	
	77	
FRANCE, US, JAPAN, CANADA	76	
FINLAND, AUSTRIA, UK	75	
USSR, GERMANY	74	
	73	
	72	SWEDEN
	71	JAPAN
	70	
	69	CANADA, UK, US, FRANCE
	68	GERMANY, AUSTRIA
	67	FINLAND
	66	
	65	
	64	
	63	USSR
	62	

Fig. 11b

Si certains auteurs peuvent masquer ou déformer des informations de manière inconsciente en utilisant de mauvaises règles graphiques, d'autres n'hésitent pas à choisir volontairement des représentations qui soulignent de manière exagérée certains traits qu'ils trouvent avantageux. Le lecteur doit ainsi rester très vigilant et toujours conscient des limites des techniques graphiques :

- L'auteur a-t-il exagéré certains traits des données ?

- La légende est-elle assez claire et assez explicite pour permettre une bonne compréhension du graphique ?

- Le graphique a-t-il conservé les données dans leur contexte ?

exemple n°12 (Tufte, 1983) :

Tufte donne l'exemple du statisticien W. Playfair qui mettait en avant les "folies ruineuses" du gouvernement anglais qu'il critiquait, en jouant sur les échelles des graphiques et en provoquant des "montées en flèches" (fig.12a). La figure n°12b présente une échelle un peu plus raisonnable que la précédente.

Fig. 12a

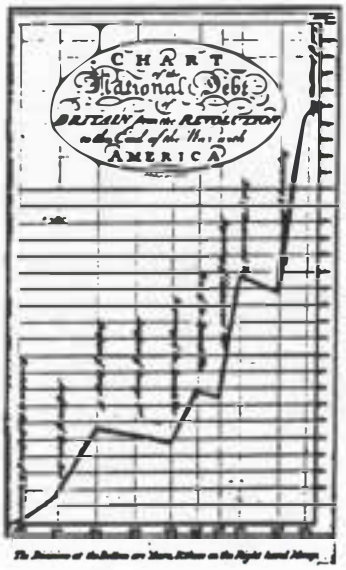
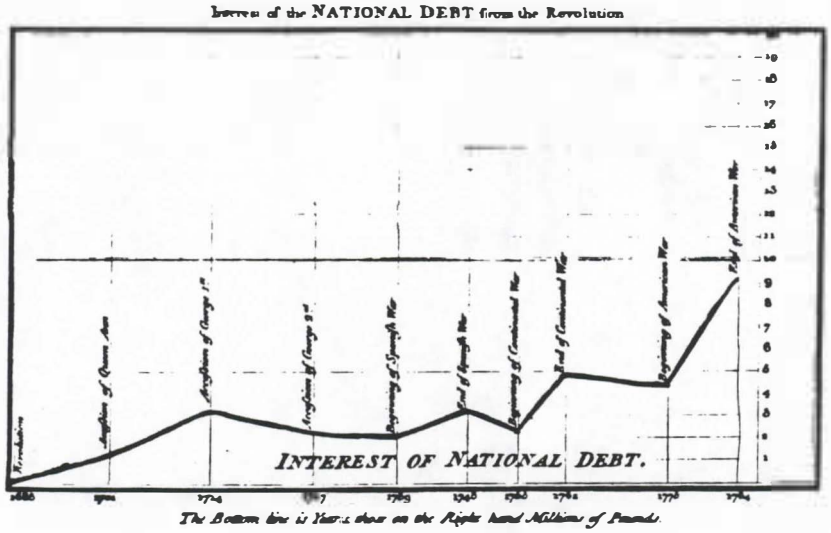


Fig. 12b



exemple n°13 (Tufte, 1983) :

Les figures n°13a et n°13b présentent un autre exemple d'exagération de certains aspects des données obtenue à l'aide d'astuces graphiques : par effets de perspectives, ces graphiques amplifient l'augmentation du budget total des trois dernières années. En outre, on peut remarquer que celles-ci ne sont pas représentées sur le même plan que les années précédentes, elles sont projetées en avant. La figure n°13c présente de manière beaucoup plus calme l'évolution du budget étudié...

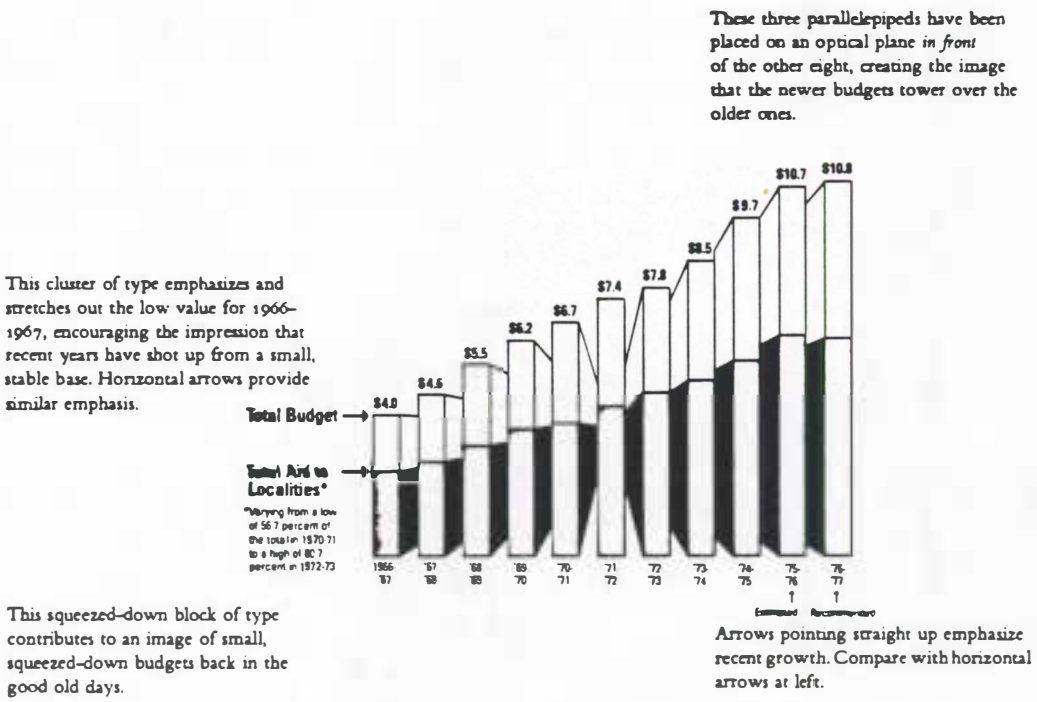


Fig. 13a

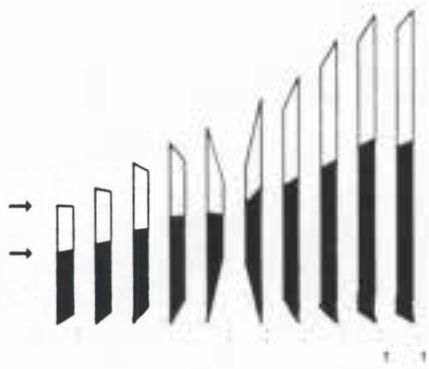


Fig. 13b

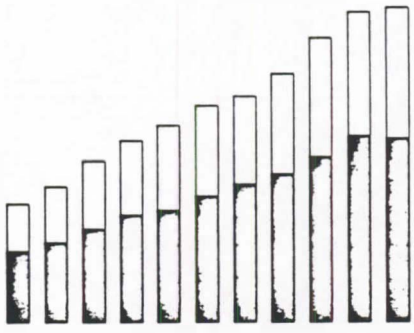


Fig. 13c

exemple n°14 (Wainer, 1984) :

Une légende insuffisante est souvent source de mauvaise compréhension du graphique et de contresens de la part du lecteur. Les auteurs du graphique de la figure n°14a avaient dans l'idée d'insister sur une soi-disant baisse des commissions touchant les agences de voyages. Un lecteur peu attentif aura de grandes chances d'abonder dans ce sens. Pourtant, une lecture plus poussée nous fait prendre conscience que la petite barre grisée, signe du "déclin", ne concerne qu'une demi-année alors que les barres précédentes concernent des années complètes. La barre "1978" oublie ainsi une grande partie des dépenses effectuées lors de périodes favorables aux voyages (*Labor Day*, *Thanksgiving*, Noël, etc.). Combien de lecteurs auront fait attention au petit "first half 78" du graphique ? En doublant la valeur de la demi-année 1978, Wainer obtient pourtant des résultats plus avantageux que ceux des années précédentes.

Fig. 14a

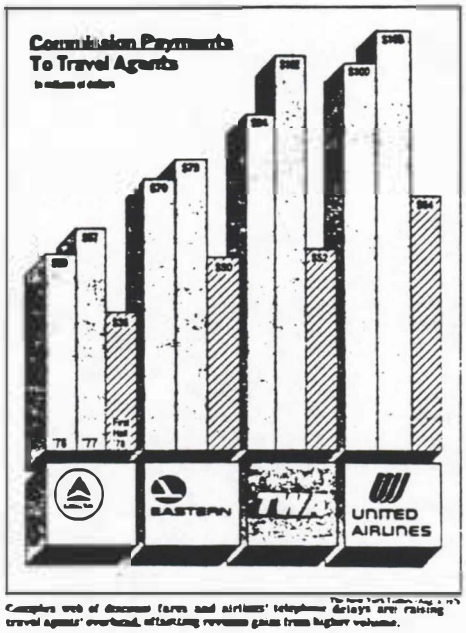
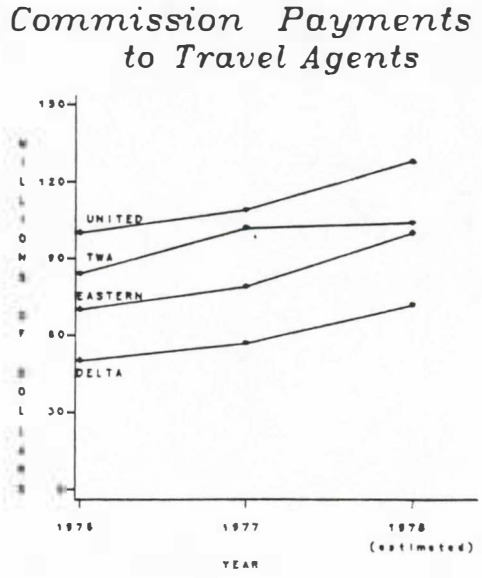


Fig. 14b



exemple n°15 (Tufte, 1983) :

Selon Tufte, le fait que les données représentées soient convenablement situées dans leur contexte est un caractère essentiel pour l'intégrité d'un graphique. Les représentations trop "minces", avec trop peu de points, doivent provoquer une réaction de suspicion chez le lecteur.

Que représente la chute des accidents routiers mortels entre 1955 et 1956 (fig.15a) par rapport à celle de la période 1955-1959 (fig.5b) ? Pour illustrer l'importance du contexte, on peut aussi imaginer les interprétations très différentes associées aux trois schémas fictifs de la figure n°15c, qui encadrent la même chute lors de la période 1955-1956.

Fig. 15a

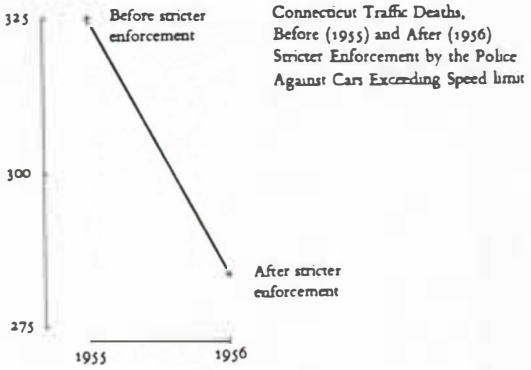


Fig. 15b

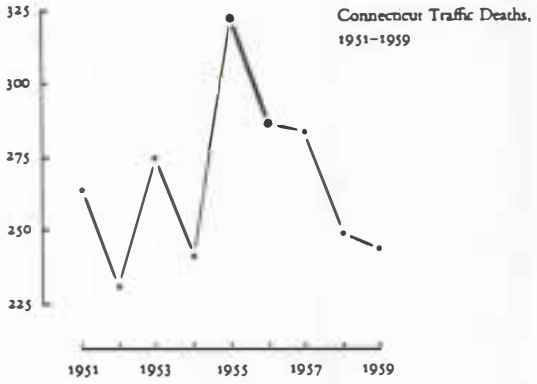
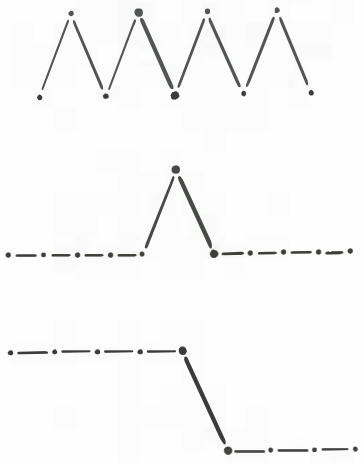


Fig. 15c



BIBLIOGRAPHIE.

- TUFTE (E.R.), 1983. The visual display of quantitative information. Cheshire, Conn., Graphic Press, 197 p.

- WAINER (H.), 1984. How to display data badly. The American Statistician, 38(2): 137-147.

LES SPLINES

Septembre 1992

Hervé LEDOUX

BIOMETRIE
CIRAD - Forêt

Les Splines

Les Splines

1. Introduction :

Le mot anglais "Spline" signifie une latte flexible utilisée par les dessinateurs pour matérialiser des lignes à courbure variable et passant par des points fixés a priori ou à "proximité" de ceux-ci. Le tracé ainsi réalisé minimise l'énergie de déformation de la latte. Par analogie, ce mot désigne également des familles de fonctions d'interpolation ou de lissage présentant des propriétés "optimales" de régularité.

Les fonctions splines d'interpolation sont des fonctions qui passent exactement par les points étudiés, tandis que les fonctions splines de lissage sont des fonctions qui passent "au plus près" des points.

Nous présenterons deux types de spline d'interpolation (chapitres 2 et 4), avec leur exemple (chapitres 3 et 5), puis nous généraliserons les splines d'interpolation (chapitre 6). Les splines de lissage seront décrites au chapitre 7 et un exemple sera présenté au chapitre 8. Nous terminerons par la présentation des possibilités de dessiner des fonctions splines avec le logiciel statistique SAS® (Statistical Analysis System).

2. Spline cubique d'interpolation :

Considérons n points (x_i , y_i) $i = 1,2,...,n$ dans le plan. Le problème d'interpolation est d'ajuster une "courbe" qui passe exactement par ces n points. Une possibilité de telle courbe est la *Spline cubique d'interpolation* s'écrivant s(x), qui est une fonction dont les dérivées sont continues jusqu'à l'ordre 2 et qui coïncide avec un polynôme de degré 3 dans chaque intervalle $[x_i, x_{i+1}]$, $i = 1,2,...n-1$.

Nous présentons le calcul des coefficients des (n-1) polynômes, constituant la fonction spline cubique d'interpolation :

Le polynôme de degré 3 d'interpolation entre les points (x_i,y_i) et (x_{i+1},y_{i+1}) s'écrit :

$$y = a_i(x-x_i)^3 + b_i(x-x_i)^2 + c_i(x-x_i) + d_i \qquad (1)$$

Les dérivées s'écrivent :

$$\frac{\delta y}{\delta x} = 3a_i(x-x_i)^2+2b_i(x-x_i)+c_i$$
$$\frac{\delta^2 y}{\delta x^2} = 6a_i(x-x_i)+2b_i$$

D'après la définition de la fonction spline, le polynôme doit :

. passer par le point (x_i, y_i) :
$$y_i = d_i$$

. passer par le point (x_{i+1}, y_{i+1}) :
$$y_{i+1} = a_i(x_{i+1}-x_i)^3 + b_i(x_{i+1}-x_i)^2 + c_i(x_{i+1}-x_i) + d_i$$

. la dérivée seconde au point x_i est égale à M_i (par définition) :
$$M_i = 2b_i$$

. la dérivée seconde au point x_{i+1} est égale à M_{i+1} (par continuité) :
$$M_{i+1} = 6a_i(x_{i+1}-x_i) + M_i$$

En résolvant ces quatre équations, nous obtenons :

$$\begin{aligned} a_i &= \frac{(M_{i+1}-M_i)}{6h_i} \\ b_i &= \frac{M_i}{2} \\ c_i &= \frac{y_{i+1}-y_i}{h_i} - \frac{2h_iM_i+h_iM_{i+1}}{6} \\ d_i &= y_i \end{aligned} \qquad (2)$$

en posant : $h_i = x_{i+1} - x_i$

Le problème d'ajustement de courbe est réduit à la recherche des valeurs de M_i. En utilisant la condition de continuité de la première dérivée de la spline, nous trouvons :

$$h_{i-1}M_{i-1}+2(h_{i-1}+h_i)M_i+h_iM_{i+1} = 6\left(\frac{y_{i+1}-y_i}{h_i} - \frac{y_i-y_{i-1}}{h_{i-1}}\right)$$

Ce système d'équation peut s'écrire sous forme matricielle :

$$\begin{vmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ h_1 & 2(h_1+h_2) & h_2 & 0 & \dots & 0 & 0 & 0 \\ 0 & h_2 & 2(h_2+h_3) & h_3 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_{n-1} \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 \end{vmatrix} * \begin{vmatrix} M_1 \\ M_2 \\ M_3 \\ \vdots \\ M_{n-1} \\ M_n \end{vmatrix} = \begin{vmatrix} O \\ 6(\frac{y_3-y_2}{h_2} - \frac{y_2-y_1}{h_1}) \\ 6(\frac{y_4-y_3}{h_3} - \frac{y_3-y_2}{h_2}) \\ \vdots \\ 6(\frac{y_n-y_{n-1}}{h_{n-1}} - \frac{y_{n-1}-y_{n-2}}{h_{n-2}}) \\ 0 \end{vmatrix}$$

Soit en simplifiant les notations :

$$U * M = L \qquad (3)$$

Ainsi, en résolvant cette équation matricielle, nous trouvons les valeurs des M_i, que nous pouvons remplacer dans les équations (2) pour obtenir les coefficients des polynômes.

3. Un exemple de spline cubique d'interpolation :

Soit les 6 points (x_i, y_i) :

i	x_i	y_i
1	1	7
2	3	3
3	6	1
4	10	3
5	12	7
6	16	9

D'après l'équation (3), nous obtenons l'équation matricielle suivante :

$$\begin{vmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 10 & 3 & 0 & 0 & 0 \\ 0 & 3 & 14 & 4 & 0 & 0 \\ 0 & 0 & 4 & 12 & 2 & 0 \\ 0 & 0 & 0 & 2 & 12 & 4 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{vmatrix} * \begin{vmatrix} M_1 \\ M_2 \\ M_3 \\ M_4 \\ M_5 \\ M_6 \end{vmatrix} = \begin{vmatrix} 0 \\ 8 \\ 7 \\ 9 \\ -9 \\ 0 \end{vmatrix}$$

L'inverse de la matrice U est égale à :

$$\begin{vmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ -0{,}21535 & 0{,}10767 & -0{,}02558 & 0{,}00877 & -0{,}00146 & 0{,}00585 \\ 0{,}05116 & -0{,}02558 & 0{,}08526 & -0{,}02923 & 0{,}00487 & -0{,}01948 \\ -0{,}01754 & 0{,}00877 & -0{,}02923 & 0{,}09573 & -0{,}01595 & 0{,}06382 \\ 0{,}00292 & -0{,}00146 & 0{,}00487 & -0{,}01595 & 0{,}08599 & -0{,}34397 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{vmatrix}$$

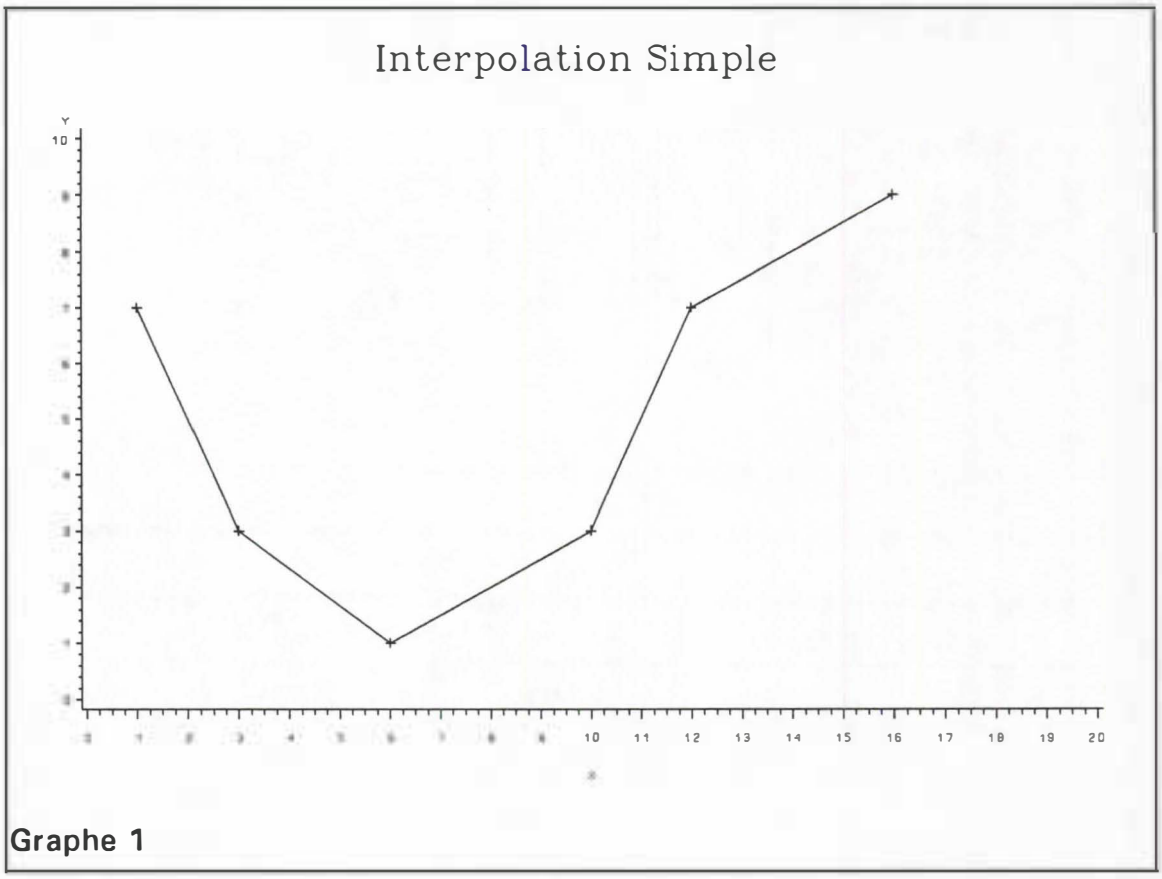
Les valeurs de M_i sont égales à :

$$\begin{vmatrix} M_1 \\ M_2 \\ M_3 \\ M_4 \\ M_5 \\ M_6 \end{vmatrix} = \begin{vmatrix} 0 \\ 0,7744 \\ 0,0853 \\ 0,8708 \\ -0,8951 \\ 0 \end{vmatrix}$$

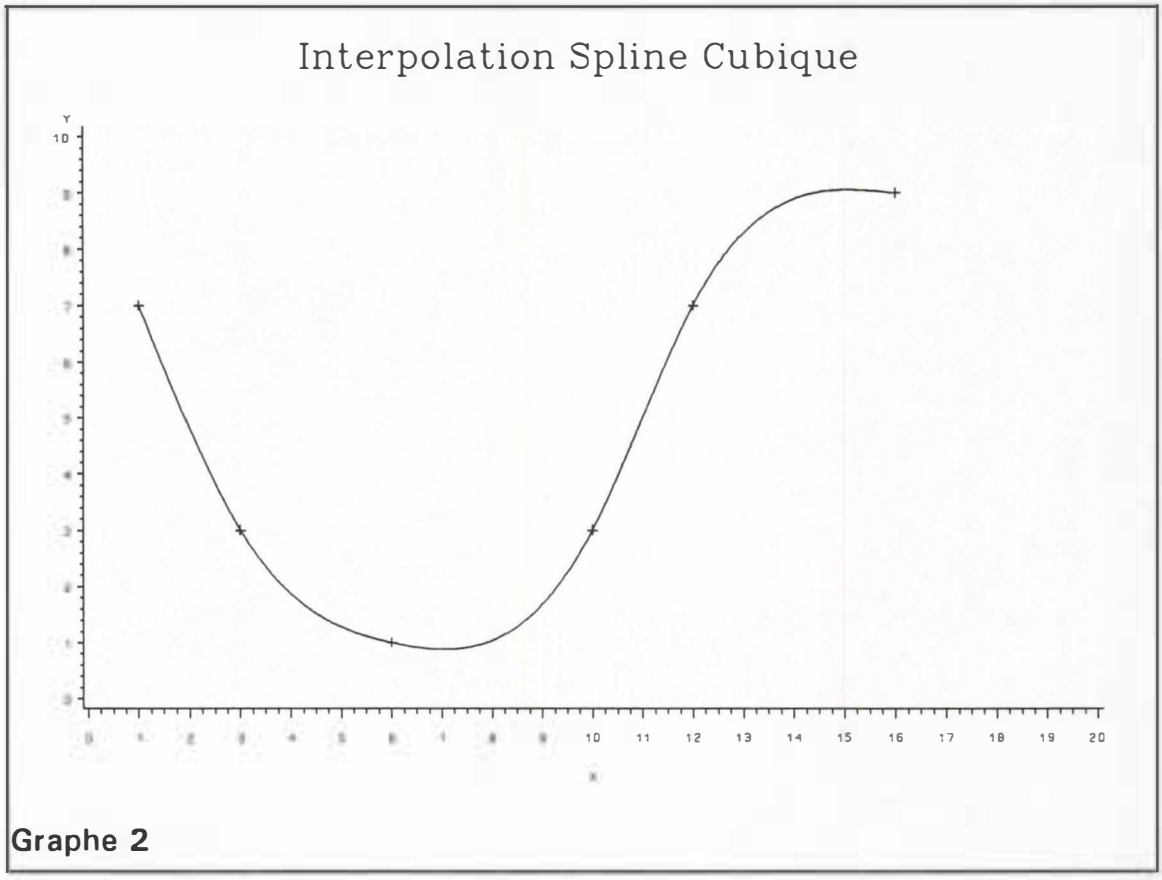
Les valeurs des coefficients des polynômes sont égaux à :

i	a_i	b_i	c_i	d_i
1	0,0645	0	-2,2581	7
2	-0,0383	0,3872	-1,4837	3
3	0,0327	0,0426	-0,1942	1
4	-0,1472	0,4354	1,7179	3
5	0,0373	-0,4476	1,6935	7

Les polynômes sont tracés sur le graphe n° 2, le graphe n° 1 représentant une interpolation des points par des droites pour une comparaison des résultats.



Graphe 1



Graphe 2

4. Spline cubique périodique d'interpolation :

Considérons n points P_i, repérés par leurs coordonnées polaires (θ_i, r_i), et répartis de façon systématique autour du centre, aux angles $(i-1)*2\pi/n$.

Une fonction spline cubique périodique d'interpolation, est un ensemble de n polynômes de degré trois au plus définis entre les points (θ_i, r_i) et (θ_{i+1}, r_{i+1}) passant par les points (θ_i, r_i) et dont les dérivées premières et deuxièmes sont continues.

Calculons les coefficients de ces polynômes :

Le polynôme de degré 3 (en θ) entre les points (θ_i, r_i) et (θ_{i+1}, r_{i+1}) s'écrit :

$$r = \alpha_i(\theta-\theta_i)^3 + \beta_i(\theta-\theta_i)^2 + \gamma_i(\theta-\theta_i) + \delta_i \qquad (4)$$

Les dérivées s'écrivent :

$$\frac{\delta r}{\delta \theta} = 3\alpha_i(\theta-\theta_i)^2 + 2\beta_i(\theta-\theta_i) + \gamma_i$$

$$\frac{\delta^2 r}{\delta \theta^2} = 6\alpha_i(\theta-\theta_i) + 2\beta_i$$

D'après la définition de la fonction spline, nous avons :

. le polynôme passe par le point (θ_i, r_i) :

$$r_i = \delta_i$$

. le polynôme passe par le point (θ_{i+1}, r_{i+1}) :

$$r_{i+1} = \alpha_i(\theta_{i+1}-\theta_i)^3 + \beta_i(\theta_{i+1}-\theta_i)^2 + \gamma_i(\theta_{i+1}-\theta_i) + \delta_i$$

. la dérivée seconde au point (θ_i, r_i) est égale M_i (par définition) :

$$M_i = 2\beta_i$$

. la dérivée seconde au point (θ_{i+1}, r_{i+1}) est égale à M_{i+1} (par continuité) :

$$M_{i+1} = 6\alpha_i(\theta_{i+1}-\theta_i) + M_i$$

. la dérivée première est continue au point (θ_i, r_i) :

$$3\alpha_{i-1}(\theta_i-\theta_{i-1})^2 + 2\beta_{i-1}(\theta_i-\theta_{i-1}) + \gamma_{i-1} = \gamma_i$$

En résolvant les quatre premières équations, nous obtenons :

$$\begin{aligned} \alpha_i &= \frac{(M_{i+1}-M_i)}{6h} \\ \beta_i &= \frac{M_i}{2} \\ \gamma_i &= \frac{r_{i+1}-r_i}{h_i} - \frac{h(2M_i+M_{i+1})}{6} \\ \delta_i &= r_i \end{aligned} \qquad (5)$$

La recherche des éléments du polynôme est réduite à la recherche des M_i, qui s'obtiennent grâce à la cinquième équation :

$$\frac{1}{2}M_{i-1}+2M_i+\frac{1}{2}M_{i+1} = \frac{6}{2h^2}(r_{i-1}-2r_i+r_{i+1})$$

Ce système d'équation peut s'écrire sous forme matricielle :

$$\begin{vmatrix} 2 & \frac{1}{2} & 0 & 0 & \dots & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 2 & \frac{1}{2} & 0 & \dots & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 2 & \frac{1}{2} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \frac{1}{2} & 2 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \dots & 0 & \frac{1}{2} & 2 \end{vmatrix} * \begin{vmatrix} M_1 \\ M_2 \\ M_3 \\ \vdots \\ M_{n-1} \\ M_n \end{vmatrix} = \frac{6}{2h^2} \begin{vmatrix} r_n-2r_1+r_2 \\ r_1-2r_2+r_3 \\ r_2-2r_3+r_4 \\ \vdots \\ r_{n-2}-2r_{n-1}+r_n \\ r_{n-1}-2r_n+r_1 \end{vmatrix}$$

Ainsi, en résolvant cette équation matricielle, nous trouvons les valeurs des M_i, que nous pouvons remplacer dans les équations (5) pour obtenir les coefficients des polynômes.

5. Un exemple de spline cubique périodique :

L'exemple présenté dans ce paragraphe provient de l'étude de la distorsion des cernes de Pin à Madagascar. Nous présenterons l'analyse d'une section de tige, les rayons étant la distance du centre théorique de la tige au cerne étudié. Cette étude a permis de calculer la surface de la section, ainsi que son évolution dans le temps (avec plusieurs mesures de cernes).

Soit les seize points :

i	r_i	θ_i
1	56	0
2	52	0,3927
3	51	0,7854
4	50	1,1781
5	48	1,5708
6	49	1,9635
7	51	2,3562
8	50	2,7489
9	51	3,1416
10	50	3,5343
11	52	3,9270
12	54	4,3197
13	54	4,7124
14	54	5,1051
15	55	5,4978
16	57	5,8905

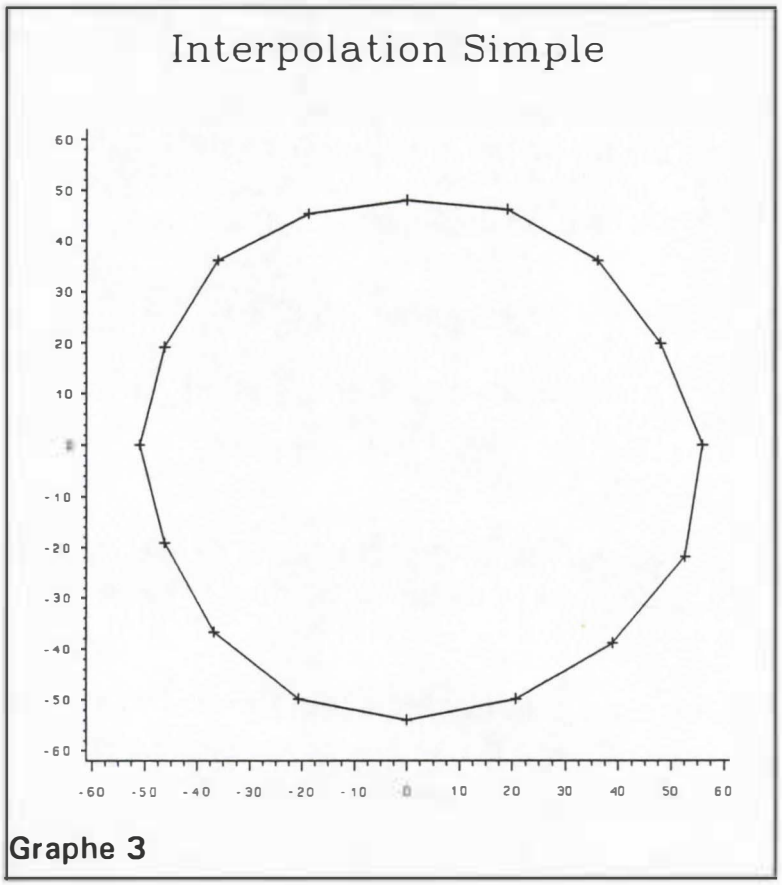
L'équation matricielle s'écrit :

$$\begin{vmatrix} 2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 \\ 1/2 & 2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/2 & 2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 2 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 2 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 2 & 1/2 \\ 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 2 \end{vmatrix} \cdot \begin{vmatrix} M_1 \\ M_2 \\ M_3 \\ M_4 \\ M_5 \\ M_6 \\ M_7 \\ M_8 \\ M_9 \\ M_{10} \\ M_{11} \\ M_{12} \\ M_{13} \\ M_{14} \\ M_{15} \\ M_{16} \end{vmatrix} = 192/\pi^2$$

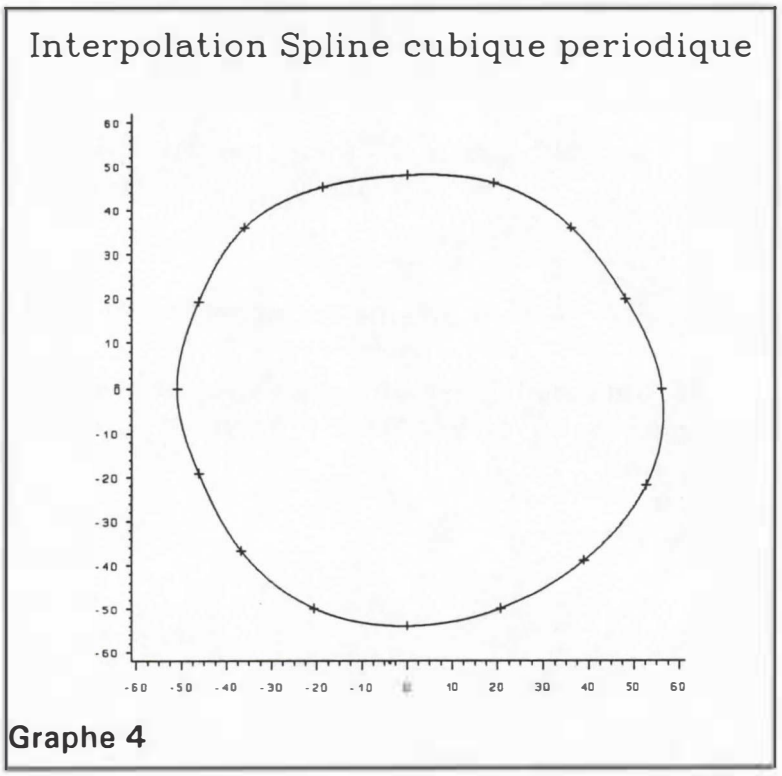
Les valeurs des coefficients des polynômes sont égaux à :

i	α_i	β_i	γ_i	δ_i
1	30,3535	-16,3594	-8,4425	56
2	-18,9113	19,3999	-7,2485	52
3	-4,2467	-2,8794	-0,7608	51
4	19,3855	-7,8824	-4,9870	50
5	-7,2441	14,9555	-2,2094	48
6	-23,4347	6,4213	6,1852	49
7	34,9316	-21,1870	0,3867	51
8	-33,7278	19,9658	-0,0928	50
9	33,9284	-19,7688	-0,0155	51
10	-19,4220	20,2021	0,1547	50
11	-5,7789	-2,6788	7,0361	52
12	9,5119	-9,4869	2,2586	54
13	0,7569	1,7190	-0,0792	54
14	3,9732	2,6108	0,9085	54
15	-16,6496	7,2915	4,7971	55
16	-3,4259	-12,3233	2,8212	57

Les polynômes sont tracés sur le graphe n° 4, le graphe n° 3 représentant une interpolation des points par des droites pour une comparaison des résultats.



Graphe 3



Graphe 4

6. Généralisation de la spline d'interpolation :

Ce paragraphe veut montrer que le concept de la spline d'interpolation ne se réduit pas à la simple utilisation d'un ensemble de polynômes.

La spline cubique d'interpolation est une solution du problème :

$$\text{Minimiser} \quad \int_{-\infty}^{+\infty} (D^2 f(x))^2 dx$$

$$\text{sachant que :} \quad D^j f \in L_2\ ]-\infty,+\infty[\qquad j=0,1,2$$
$$\text{et :} \quad f(x_i) = y_i \qquad i=1,2,...,n$$

où D est l'opérateur différentiel (ainsi D^2 est la dérivée seconde de la fonction f), et L_2 l'ensemble des fonctions dont le carré est intégrable.

Ce problème peut être généralisé. Considérons L comme un opérateur différentiel d'ordre m avec coefficient constant ($L = \sum_{i=1}^{m} a_i D^i$) , l'équation précédente peut s'écrire :

$$\text{Minimiser} \quad \int_{-\infty}^{+\infty} (Lf(x))^2 dx$$

$$\text{sachant que :} \quad D^j f \in L_2\ ]-\infty,+\infty[\qquad j=0,1,...,m$$
$$\text{et :} \quad f(x_i) = y_i \qquad i=1,2,...,n$$

La solution s(x) de ce problème est appelée une *L spline d'interpolation*.

Si $L = D^m$, alors s(x) est un polynôme par morceau de degré 2m-1. Si $L = D^2$, la spline est un polynôme par morceau de degré 3 (*spline cubique d'interpolation*).

7. Spline de lissage :

Dans certain cas, les points que l'on désire joindre sont connus avec une certaine imprécision. Il est donc intéressant de créer une fonction lissée qui passe près de ces points. Cette fonction est une spline de lissage.

Nous présentons une des méthodes d'estimation de la spline de lissage, il s'agit de la méthode des "moindres carrés pénalisés" :

Le problème s'écrit :

$$\text{Minimiser} \quad \sum_{i=1}^{n} (f(x_i)-y_i)^2 + \lambda \int_0^1 (Lf(x))^2 dx$$

sachant que : $D^j f$ est continu pour $j=0,1,\ldots,m-1$

$D^m \in L_2\,[0,1]$

$\lambda > 0$ fixé

Le critère à minimiser dépend d'un terme des moindres carrés, et d'un terme favorisant le lissage. Ce deuxième terme étant contrôlé par le paramètre λ. Si λ est proche de zéro, le critère correspond à une régression linéaire (minimisation des moindres carrés), tandis que si λ tend vers l'infini, le terme de lissage est prépondérant et la fonction estimée est proche d'une droite (dans le cas d'une spline cubique de lissage).

Le choix du paramètre λ est très important. La méthode de validation croisée (Cross Validation) permet de sélectionner une valeur de ce paramètre. Cette méthode est basée sur le fait que si nous omettons le $k^{ème}$ point et que nous estimons une nouvelle fonction spline $s_\lambda^{(k)}$, cette fonction à l'abscisse x_k est un estimateur de l'ordonnée y_k, et λ est choisi correctement si $s_\lambda^{(k)}(x_k)$ est un "bon" estimateur de y_k. Pour mesurer la qualité de l'ajustement, nous utilisons la moyenne du carré de l'erreur :

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n} (s_\lambda^{(k)}(x_k)-y_k)^2$$

La valeur de λ est choisie pour minimiser $CV(\lambda)$.

Une autre méthode de sélection du paramètre λ a été mise au point, il s'agit de la validation croisée généralisée. Cette méthode dérive de la précédente par une pondération des points, pour tenir compte des intervalles irréguliers entre les points et d'autres effets de bord. Le paramètre λ est choisi pour minimiser $GCV(\lambda)$:

$$GCV(\lambda) = \frac{1}{n}\sum_{k=1}^{n} (s_\lambda^{(k)}(x_k) - y_k)^2 \; w_k(\lambda)$$

$$= \frac{\frac{1}{n} \| (I - A(\lambda))y \|^2}{\left(\frac{1}{n} \mathrm{Trace}(I - A(\lambda)) \right)^2}$$

où $A(\lambda)$ est une matrice de dimension n*n, tel que $s = A(\lambda) * y$.

8. Un exemple de spline cubique de lissage :

L'exemple présenté dans ce chapitre utilise la méthode d'estimation présentée par REINSCH (1967). Cette méthode d'estimation diffère dans sa présentation de celle décrite dans le chapitre précédent.

La fonction spline s'estime par :

$$\text{Minimiser} \quad \int_{x_1}^{x_n} (D^2 f(x))^2 dx$$

$$\text{sachant que :} \quad \sum_{i=1}^{n} \left(\frac{f(x_i)-y_i}{\delta y_i}\right)^2 \leq S$$

$$\text{et :} \quad D^j f \in L_2\ ]-\infty,+\infty[\quad j=0,1,2$$

où δy_i représente une pondération, et S un nombre donné.

Minimiser l'équation précédente équivaut à minimiser :

$$\int_{x_1}^{x_n} (D^2 f(x))^2 dx + p \left\{ \sum_{i=1}^{n} \left(\frac{f(x_i)-y_i}{\delta y_i}\right)^2 + z^2 - S \right\}$$

où z est une variable auxiliaire, nécessaire pour la résolution de l'équation sous contrainte par la méthode de Lagrange. La solution est un polynôme par morceau de degré 3 identique au polynôme défini à l'équation (1).

Les coefficients des polynômes s'écrivent :

$$\begin{aligned} & b_1 = b_n = 0 \\ & a_i = \frac{(b_{i+1}-b_i)}{3h_i} \qquad i=1,\ldots,n-1 \\ & c_i = \frac{d_{i+1}-d_i}{h_i} - \frac{b_i}{h_i} - \frac{a_i}{h_i^2} \qquad i=1,\ldots,n-1 \\ & T\, b = Q^T d \\ & Q\, b = p\, D^{-2} (\, y - d\,) \end{aligned} \tag{6}$$

avec les notations suivantes :

$h_i = x_{i+1} - x_i$
$b = (b_2,\ldots,b_{n-1})^T$
$y = (y_1,\ldots,y_n)^T$
$d = (d_1,\ldots,d_n)^T$
$D = \text{Diag}\ (\delta y_1,\ldots,\delta y_n)$
T est une matrice tridiagonale défini positive d'ordre (n-2)

Q est une matrice tridiagonale avec n lignes et (n-2) colonnes.

$$T = \begin{vmatrix} \frac{2(h_1+h_2)}{3} & \frac{h_2}{3} & 0 & \dots & 0 & 0 & 0 \\ \frac{h_2}{3} & \frac{2(h_2+h_3)}{3} & \frac{h_3}{3} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \frac{h_{n-2}}{3} & \frac{2(h_{n-3}+h_{n-2})}{3} & \frac{h_{n-1}}{3} \\ 0 & 0 & 0 & \dots & 0 & \frac{h_{n-1}}{3} & \frac{2(h_{n-2}+h_{n-1})}{3} \end{vmatrix}$$

$$Q = \begin{vmatrix} \frac{1}{h_1} & 0 & 0 & \dots & 0 & 0 & 0 \\ -(\frac{1}{h_1}+\frac{1}{h_2}) & \frac{1}{h_2} & 0 & \dots & 0 & 0 & 0 \\ \frac{1}{h_2} & -(\frac{1}{h_2}+\frac{1}{h_3}) & \frac{1}{h_3} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & \frac{1}{h_{n-2}} & -(\frac{1}{h_{n-2}}+\frac{1}{h_{n-1}}) \\ 0 & 0 & 0 & \dots & 0 & 0 & \frac{1}{h_{n-1}} \end{vmatrix}$$

Pour calculer ces coefficients, nous utilisons l'algorithme suivant :

Au Début : $p = 0$

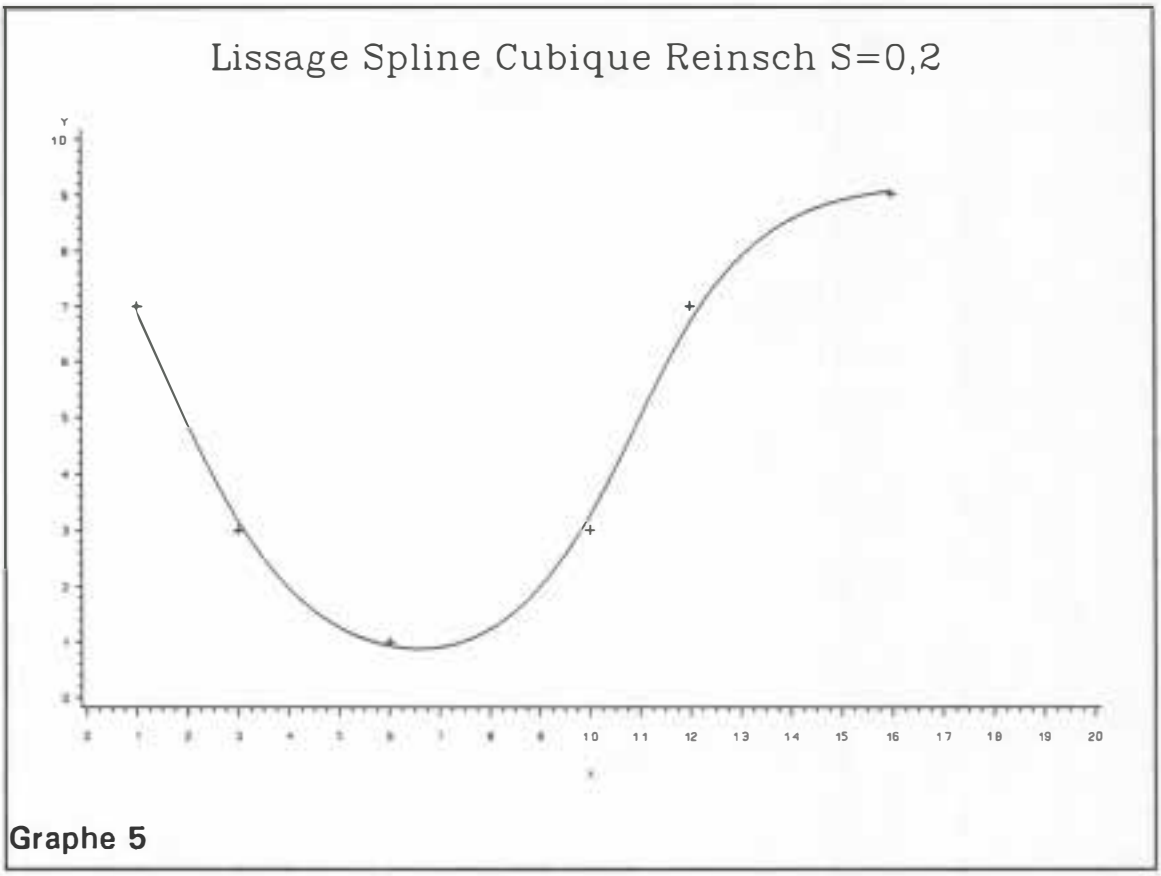
(i) Décomposition de Cholesky : $R^TR = Q^TD^2q + pT$

(ii) Calcul de u, v et e : $R^TRu = Q^Ty$
$v = DQU$
$e = v^Tv$

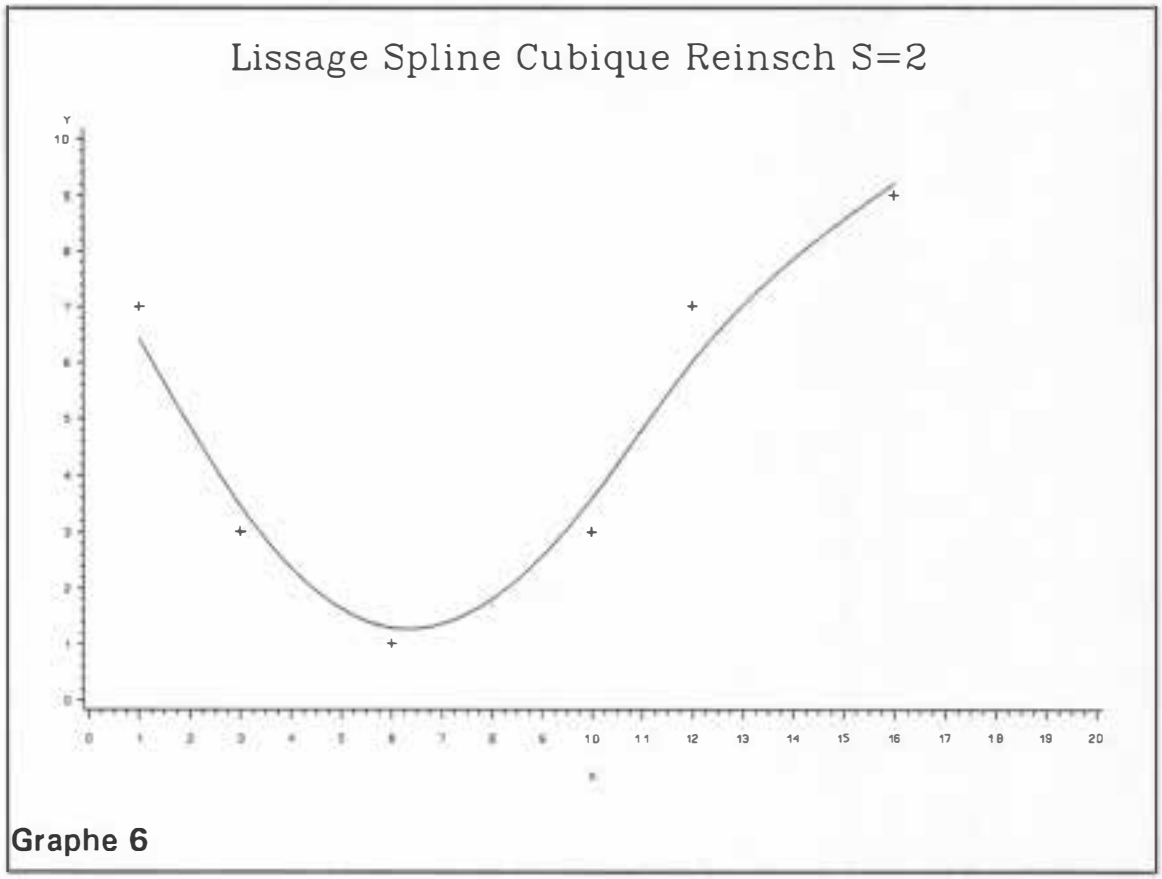
(iii) Si e est plus grand que S :
Calcul de $f = u^TTu$, $g = w^Tw$ où $R^Tw = Tu$
Remplacement de p par $p + (e-(Se)^{1/2})/(f-p*g)$
Retour à (i)

(iv) Calcul de $d = y-Dv$, $b = pu$, et a_i, c_i d'après l'équation (6).

Cette méthode a été présenté sur l'exemple des six points du paragraphe 3. Les graphes 5 et 6 présentent les résultats pour une valeur de S égale respectivement à 0,2 et 2.



Graphe 5



Graphe 6

9. SAS a le spline :

Le logiciel statistique SAS® permet de tracer facilement des fonctions splines d'interpolation ou de lissage. Dans les deux cas les fonctions utilisées par le logiciel sont des fonctions splines cubiques.

L'utilisation des fonctions splines se fait à l'aide de la procédure **GPLOT** qui permet de tracer des graphiques, et est définie dans l'option **INTERPOL** de la commande **SYMBOL**.

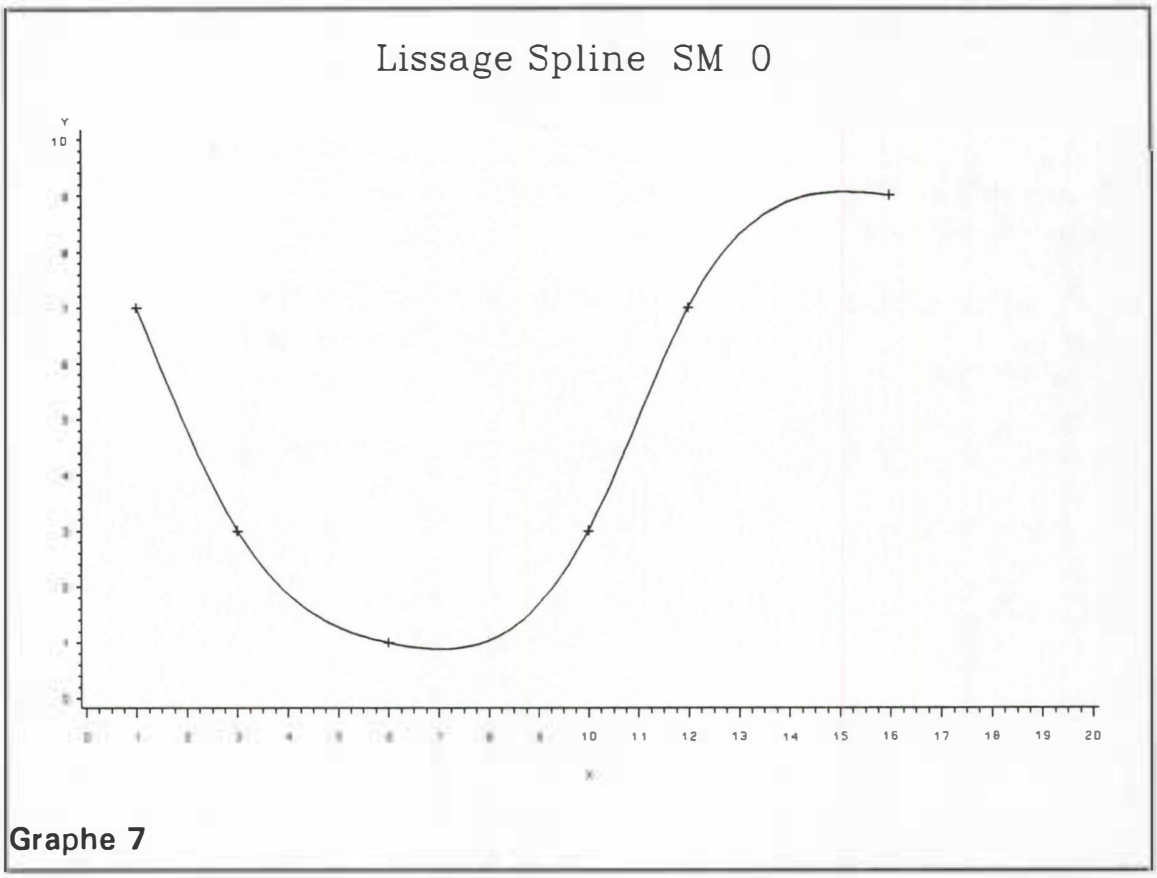
Les valeurs prises par l'option INTERPOL sont les suivantes :

INTERPOL =	JOIN	relie les points par des droites.
	SPLINE	relie les points par une spline cubique d'interpolation.
	SMnn	trace une spline de lissage qui passe "au plus près" des points. nn est une valeur comprise entre 0 et 99. Elle définit le taux de lissage. Si nn=0 la courbe est une spline cubique d'interpolation (pas de lissage), si nn=99 la courbe obtenue est presque une droite.

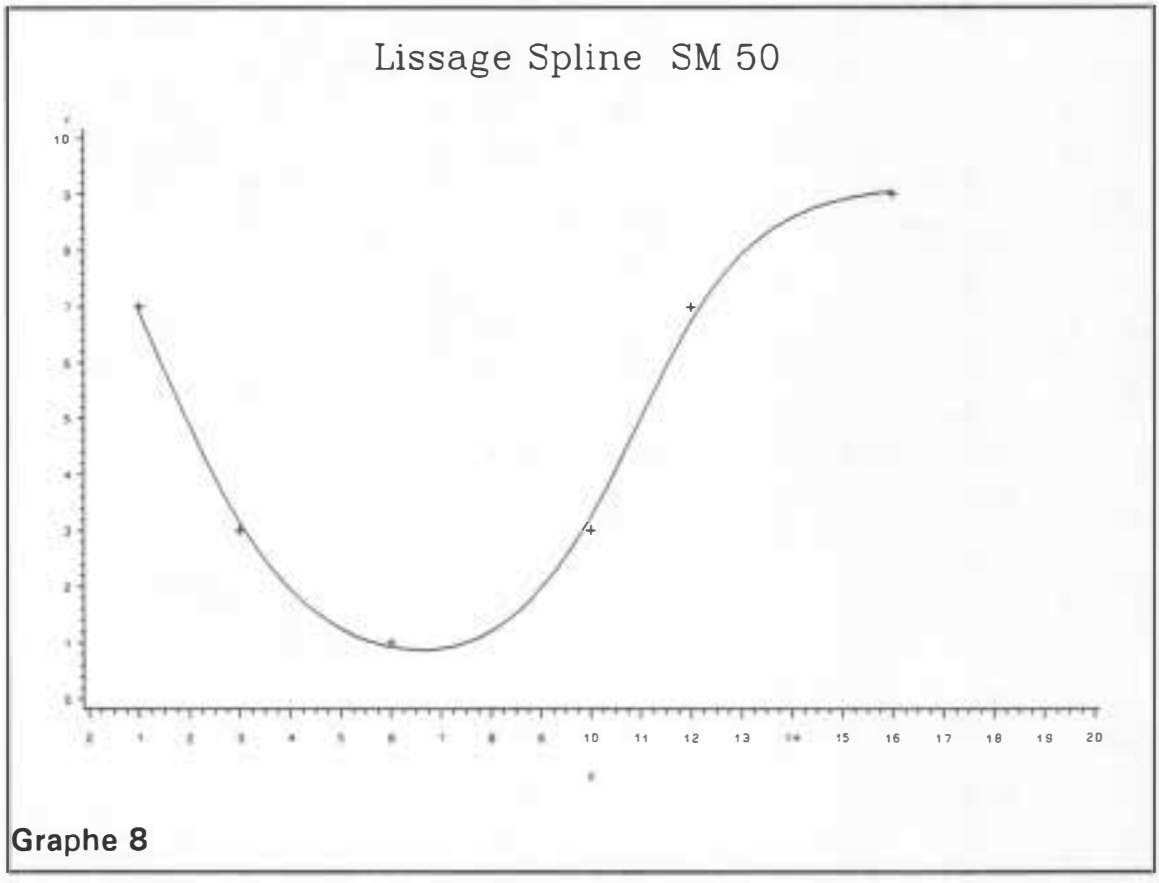
Des exemples sont présentés tout au long de la note, ainsi le graphe 1 représente l'option INTERPOL=JOIN, le graphe 2 représente l'option INTERPOL=SPLINE, et les graphes 7 à 9 représentent respectivement les options INTERPOL= SM0, SM50 et SM99

Exemple d'un programme SAS :

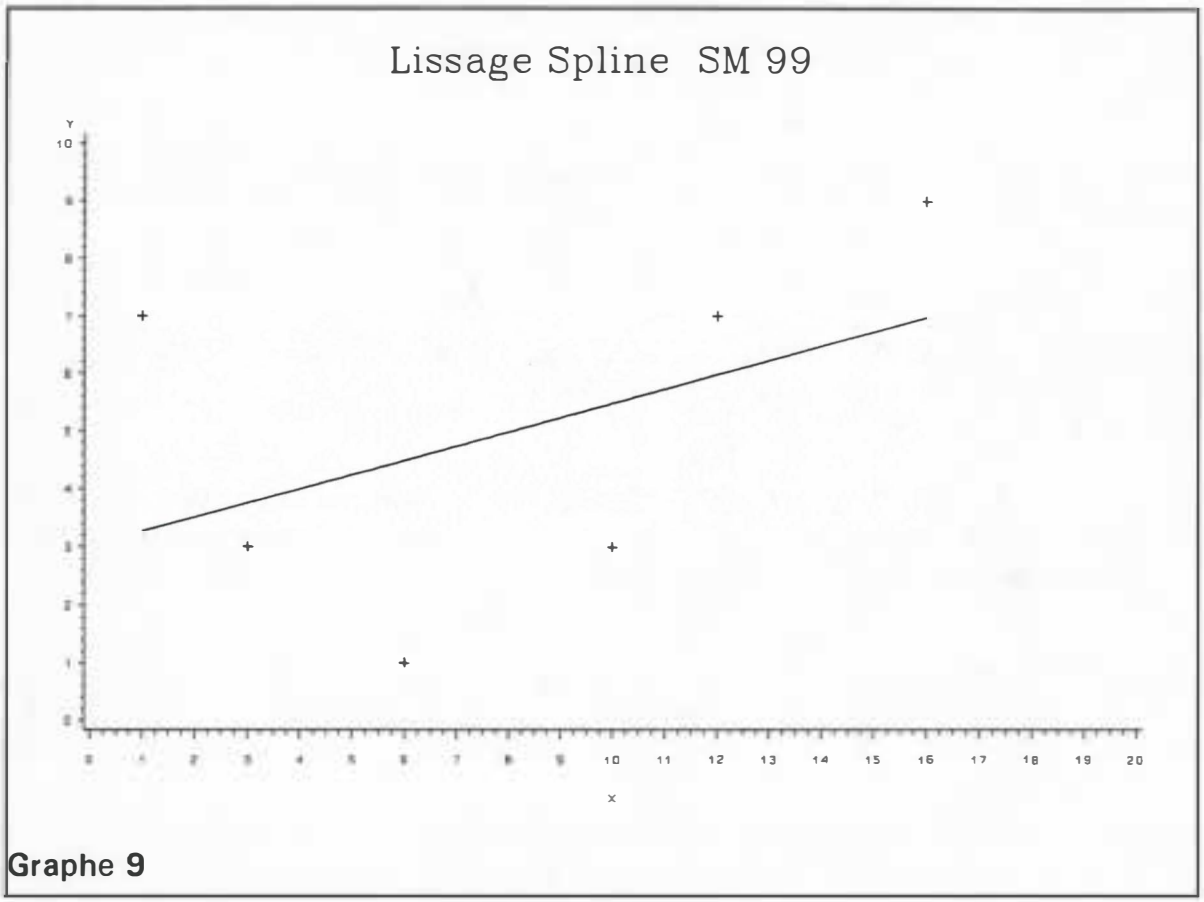
```
filename fichier 'donnees.dat';
filename sortie 'hp7475a.out';
goptions device=hp7475a gsfmode=replace gsfname=sortie
        nodosplay noprompt handshake=xonxoff
data donnee;
  infile fichier;
  input x y;

symbol1 color=blue interpol=spline value=plus;

proc gplot;
  plot y*x;
run;
```



Graphe 7



Graphe 8



Graphe 9

Références

ABBAR H. (1990), "Un estimateur spline du contour d'une répartition ponctuelle aléatoire", *Statistique et Analyse des Données*, Volume 15 n° 3, pages 1-19.

BERGONZINI J-C., BLANC N. et BOUILLET J-P. (1991), "Etude de la distorsion des cernes", Note interne C.T.F.T., 30 pages.

BESSE P. et THOMAS-AGNAN C. (1989), "Le lissage par fonctions splines en statistique. Revue bibliographique", *Statistique et Analyse des Données*, Volume 14, Numéro 1, pages 55-84.

REINSCH Christian H. (1967), "Smoothing by spline functions", *Numerische Mathematik*, February 1967, Volume 10, pages 177-183.

WEGMAN Edward J. and WRIGHT IAN W. (1983), "Splines in Statistics", *Journal of the American Statistical Association*, June 1983, Volume 78, Numéro 382, pages 351-365.

KHI-PLOTS ET INDEPENDANCE ENTRE DEUX VARIABLES

Septembre 1992

Jean-Claude BERGONZINI

BIOMETRIE
CIRAD - Forêt

KHI-PLOTS ET INDEPENDANCE ENTRE DEUX VARIABLES

1. Représentation cartésienne et Khi-deux

On considère deux variables (X,Y) et une représentation cartésienne d'un n échantillon (x_i, y_i) associé.

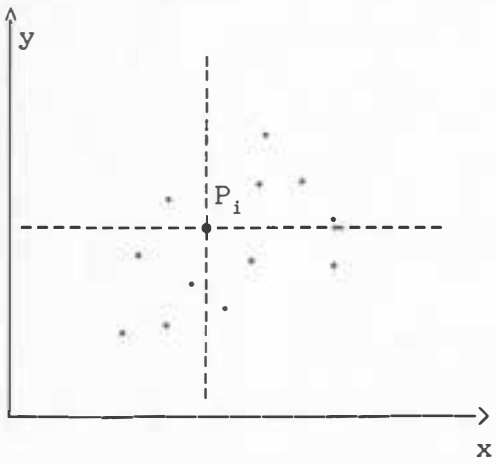
Pour chaque point P_i de coordonnées (x_i, y_i), on découpe le plan en quatre régions A_i B_i C_i et D_i :

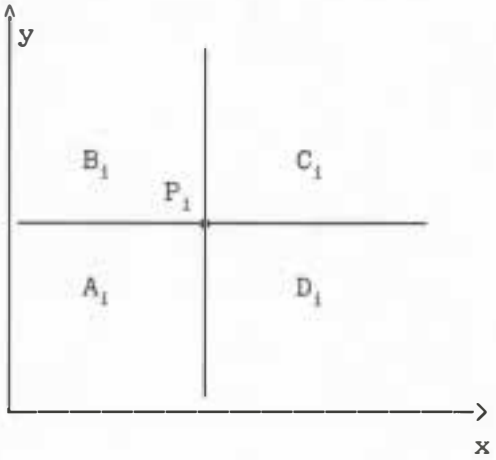
A_i = { ensemble des points $x_e \leq x_i$ et $y_e \leq y_i$ } $\quad card(A_i) = a_i$

B_i = { ensemble des points $x_e \leq x_i$ et $y_e > y_i$ } $\quad card(B_i) = b_i$

C_i = { ensemble des points $x_e > x_i$ et $y_e > y_i$ } $\quad card(C_i) = c_i$

D_i = { ensemble des points $x_e > x_i$ et $y_e \leq y_i$ } $\quad card(D_i) = d_i$





.../...

On considère le tableau suivant

Tableau T_i :

	$x_e \leq x_i$	$x_e > x_i$
$y_e > y_i$	b_i	c_i
$y_e \leq y_i$	a_i	d_i

L'indépendance de X et Y au point P_i peut être évaluée par le calcul du Khi-deux associé au tableau T_i

$$q_i^2 = \frac{(n-1)(a_i\ c_i - b_i\ d_i)^2}{(a_i + b_i)(c_i + d_i)(b_i + c_i)\ (a_i + d_i)}$$

Remarque 1 : q_i^2 est la valeur prise par un Khi-deux à un degré de liberté - autrement dit q_i est la valeur prise par une $\mathcal{N}(0,1)$.

Remarque 2 : q_i n'est défini que dans la mesure où le dénominateur est différent de zéro. De même, l'approximation par la loi normale n'est acceptable que si les effectifs estimés de chaque classe (A, B, C et D) sont suffisamment grands (en général supérieurs à 5). En fait, on élimine les points P_i pour lesquels

$$|q'_i| \geq 4\ (\frac{1}{n-1} - \frac{1}{2})^2$$

(voir définition de q'_i ci-dessous).

Dans la pratique, FISHER et SWITZER qui, les premiers, ont introduit cette méthode, proposent de calculer

$$q'_i = \frac{a_i\ c_i - b_i\ d_i}{(a_i + b_i)(c_i + d_i)(b_i + c_i)\ (a_i + d_i)}$$

qui peut être considéré comme la valeur du coefficient phi (voir *Annexe*).

q'_i est la valeur prise par $\mathcal{N}(0,\frac{1}{n})$

et $\bar{q}_i = \sin(\frac{1}{2}\ \pi\ q'_i)$ pour les représentations graphiques.

Remarque 3 : $-1 \leq q'_i \leq 1$ et $-1 \leq \bar{q}_i \leq 1$

.../...

2. Présentation de la méthode du Khi-plot

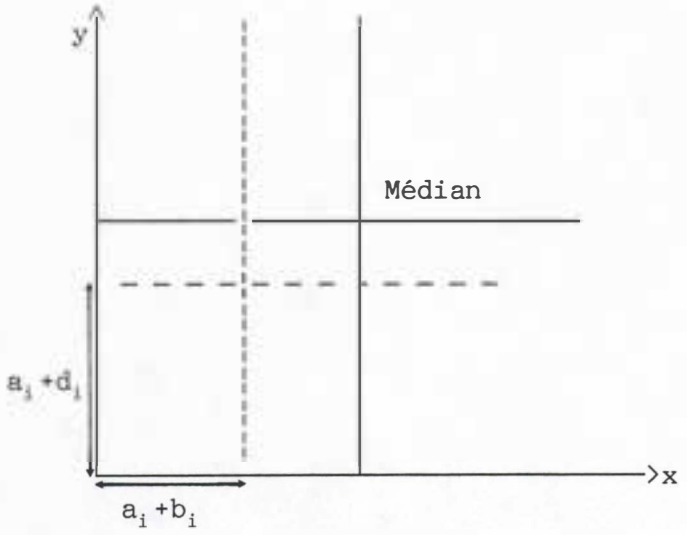
On est à la tête d'un grand nombre de valeurs q'_i ou $\bar{q}_i$. Chaque q'_i est associé à un point P_i que l'on va positionner dans le nuage par rapport au point médian.

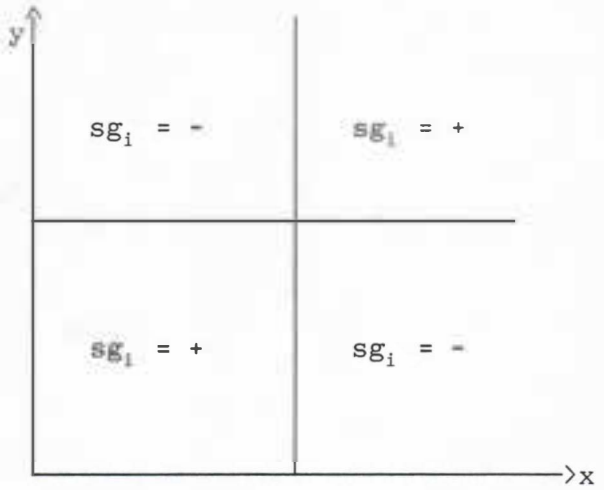
On calcule :

$$\lambda_i = 4\, sg_i \max\left[\left(\frac{a_i + b_i}{n-1} - \frac{1}{2}\right)^2 , \left(\frac{a_i + d_i}{n-1} - \frac{1}{2}\right)^2 \right]$$

$$sg_i = \text{signe de } \left(\frac{a_i + b_i}{n-1} - \frac{1}{2}\right)\left(\frac{a_i + d_i}{n-1} - \frac{1}{2}\right)$$

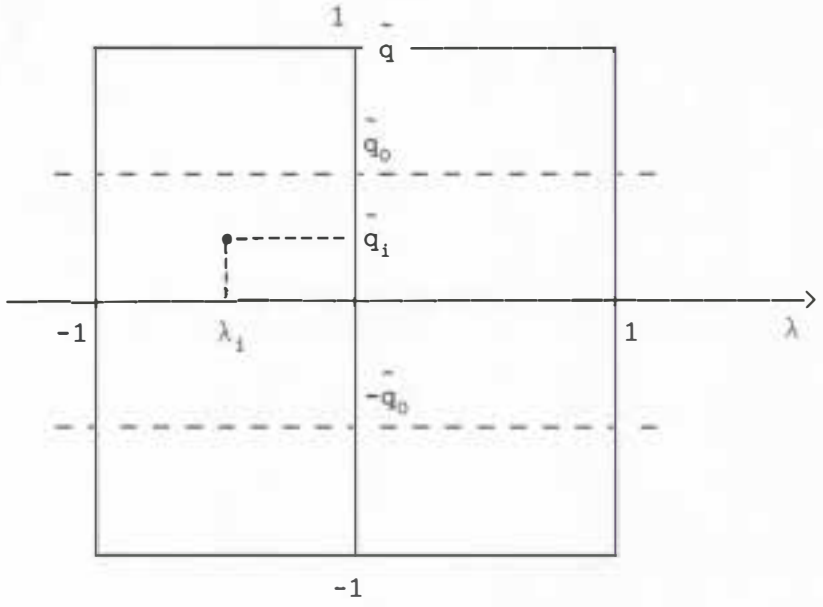
$|\lambda_i|$ est grand si le point P_i est distant du point médian dans l'une des deux directions x et/ou y .





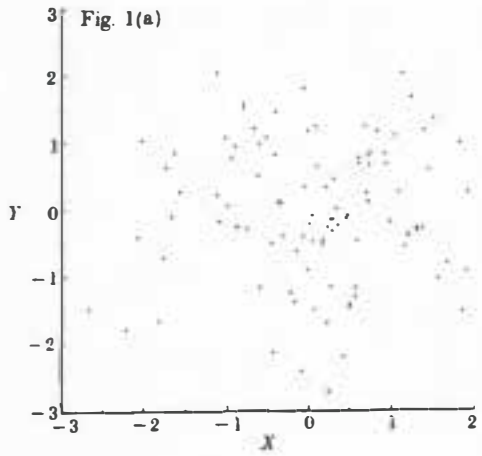
.../...

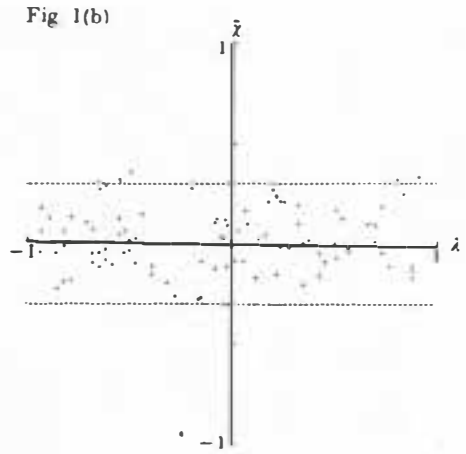
Le Khi-plot est la représentation des points $T_i = (\lambda_i , \bar{q}_i)$



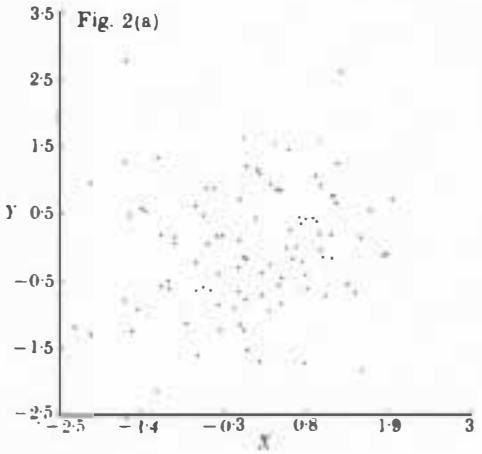
q_0 : limite pour un intervalle de confiance (0,95).

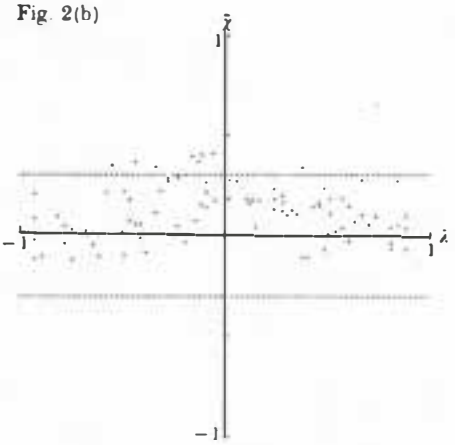
.../...



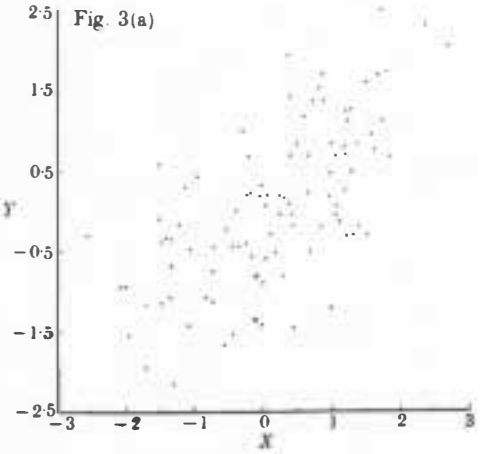


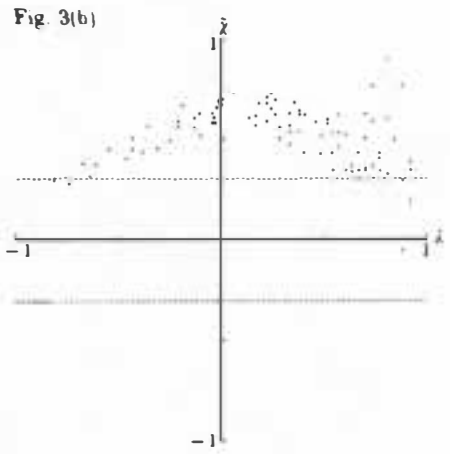
Simulation de deux variables (X,Y) normales indépendantes (ρ = 0)





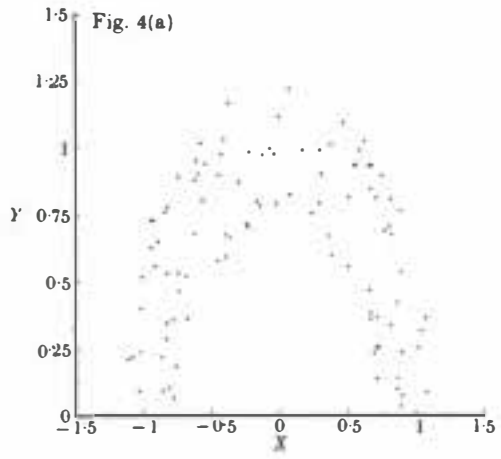
Simulation de deux variables (X,Y) normales avec ρ = 0,1

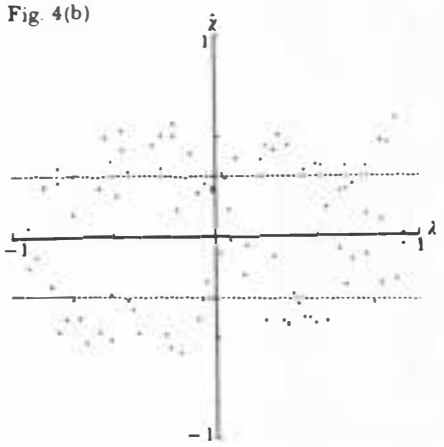




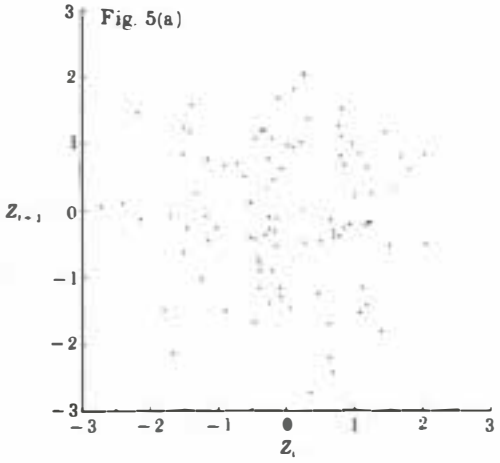
Simulation de deux variables (X,Y) normales avec ρ = 0,5

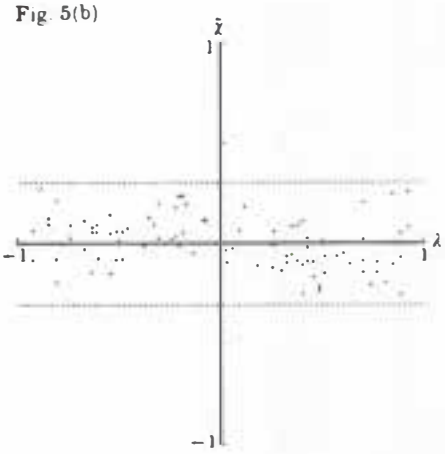
.../...



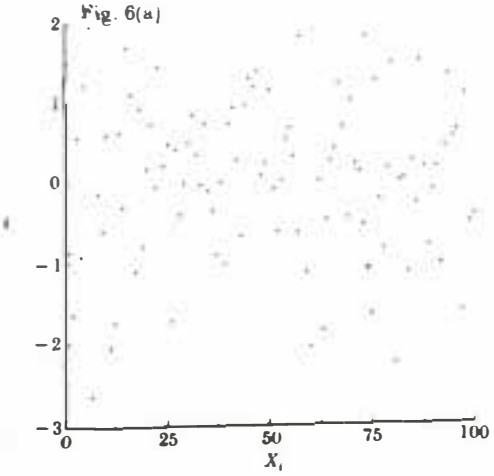


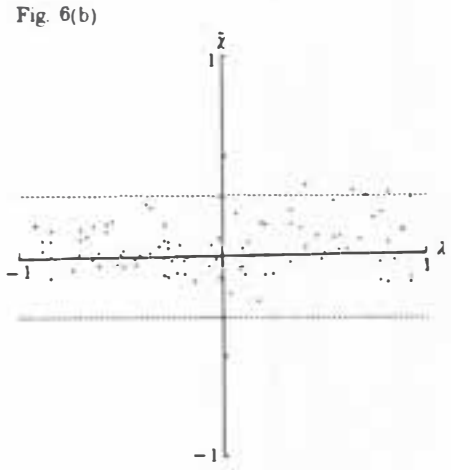
Simulation de deux variables (X,Y) normales avec $\rho = 0$ et la contrainte $Y > 0$ $\quad |X^2 + Y^2 - 1| < 1/2$





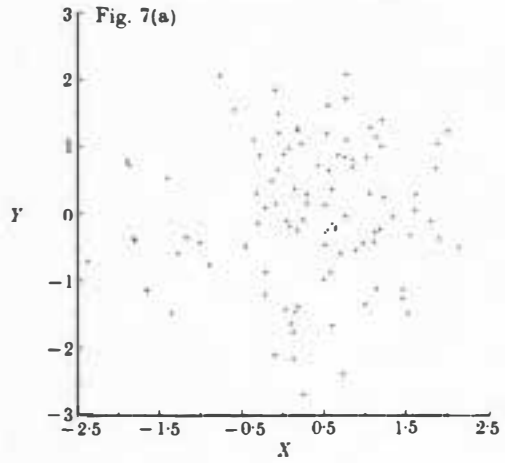
Simulation d'une variable Z de loi normale et construction des couples (Z_i , Z_{i+1}) - les tirages sont indépendants

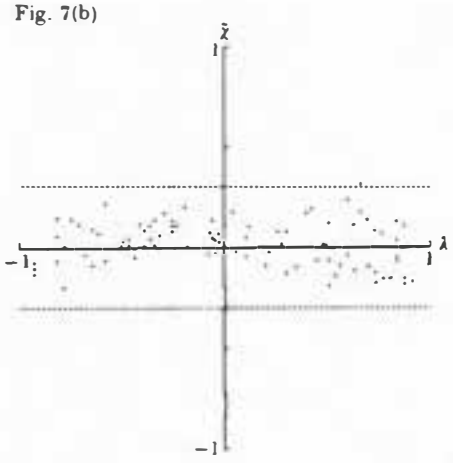




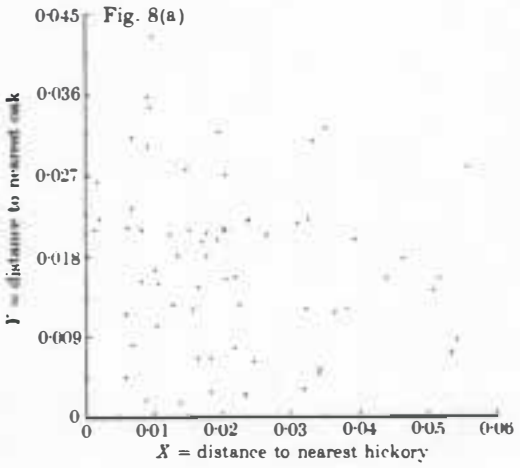
Simulation d'une variable Z de loi normale et construction des couples (Z_i , i).

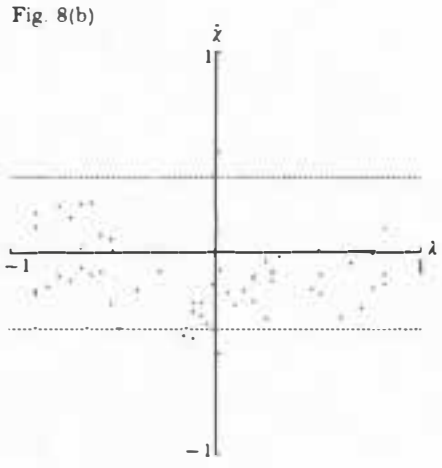
.../...



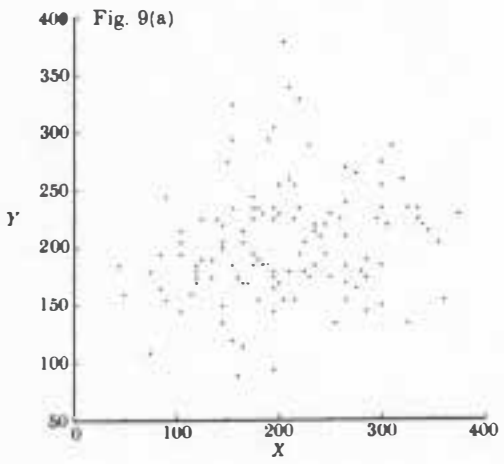


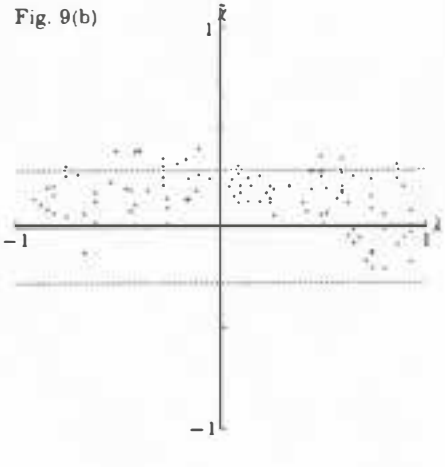
Deux variables (X,Y) de loi normale avec $\rho_{X_i X_{i+1}} = 0$ $\rho_{Y_i Y_{i+1}} = 0,9$





On considère deux espèces forestières. On choisit un point au hasard et l'on mesure X la distance à la première espèce et Y la distance à la seconde espèce.





Nuage construit en créant des liaisons en périphérie.

LE GRAPHIQUE TRIANGULAIRE

Septembre 1992

Matthieu LESNOFF

BIOMETRIE
CIRAD - Forêt

LE GRAPHIQUE TRIANGULAIRE.

Le graphique triangulaire est utilisé pour représenter un ensemble d'éléments dont une grandeur constante est fractionnée en trois grandeurs variables : le cas typique est celui de la ventilation de pourcentages en trois postes A, B, C.

On considère un triangle équilatéral dont chaque côté représente l'un des postes A, B, C. Chacun des points M inscrits dans le triangle est défini par ses trois coordonnées a, b, c dont la somme est constante : dans le cas de pourcentages a+b+c=100%. Ces coordonnées sont déterminées à l'aide des parallèles aux côtés du triangle passant par M ; il faut préalablement définir un sens de lecture (fig.1).

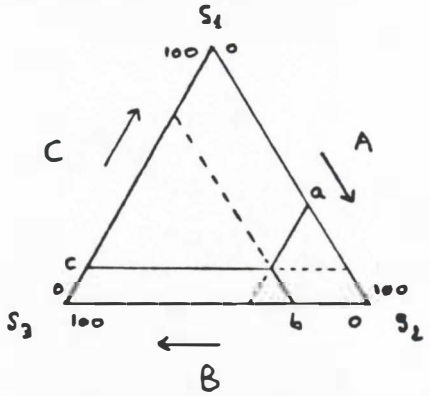


Fig.1 : Principe du graphique triangulaire. Quel que soit le point M inscrit dans le triangle, a+b+c=constante (ici 100%).

Certains éléments de la figure n°1 ont des propriétés singulières. Les côtés du triangle correspondent à l'égalité à zéro d'un des trois pourcentages (S_1S_2 : b=0), ses sommets à l'égalité à 100% d'un des pourcentages (S_1 : c=100 a=b=0). Les milieux des côtés correspondent à l'égalité à 50% de deux pourcentages (milieu de S_1S_2 : a=c=50 b=0), et le centre du triangle à l'égalité des trois pourcentages. Enfin, les hauteurs délimitent six régions dans lesquelles les trois coordonnées sont rangées dans le même ordre (fig.2).

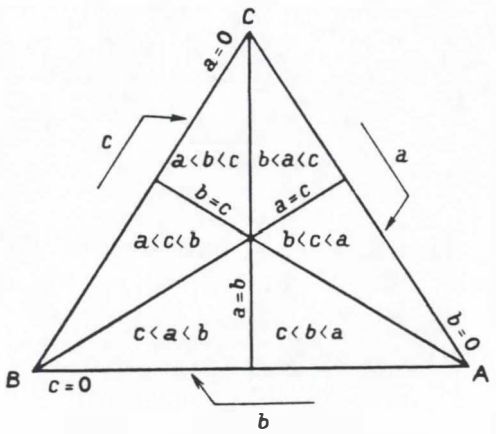


Fig.2 : Les différentes régions du graphique triangulaire (D'après Callot, 1984).

Un premier exemple (Callot, 1984) montre l'intérêt du graphique triangulaire lorsque l'on veut suivre l'évolution de trois pourcentages simultanément le long d'une échelle ordonnée (le temps par exemple). Le graphique de la figure n°3 présente l'évolution de l'état matrimonial de la population française masculine de 1960 en fonction de l'âge des individus. La population est répartie en trois postes : les célibataires, les mariés, et les veufs ou divorcés. L'évolution de ces trois postes a été suivie pour les âges allant de 18 ans à 90 ans. La courbe passant par les différents points représente le chemin matrimonial de la population.

L'ensemble des individus âgés de 18 ans sont célibataires (100%). Le pourcentage des célibataires diminue ensuite régulièrement jusqu'à l'âge de 29 ans, au profit des mariés qui atteignent alors plus de 70%. A partir de 30 ans, la baisse du pourcentage des célibataires devient le fait de l'augmentation conjointe des mariés et des veufs ou divorcés. On observe enfin, dès l'âge de 60 ans, une stabilisation des célibataires autour de 5% et une augmentation progressive des veufs ou divorcés au dépourvu des mariés. A 90 ans, les individus mariés ne représentent plus que 30% environ de la population initiale.

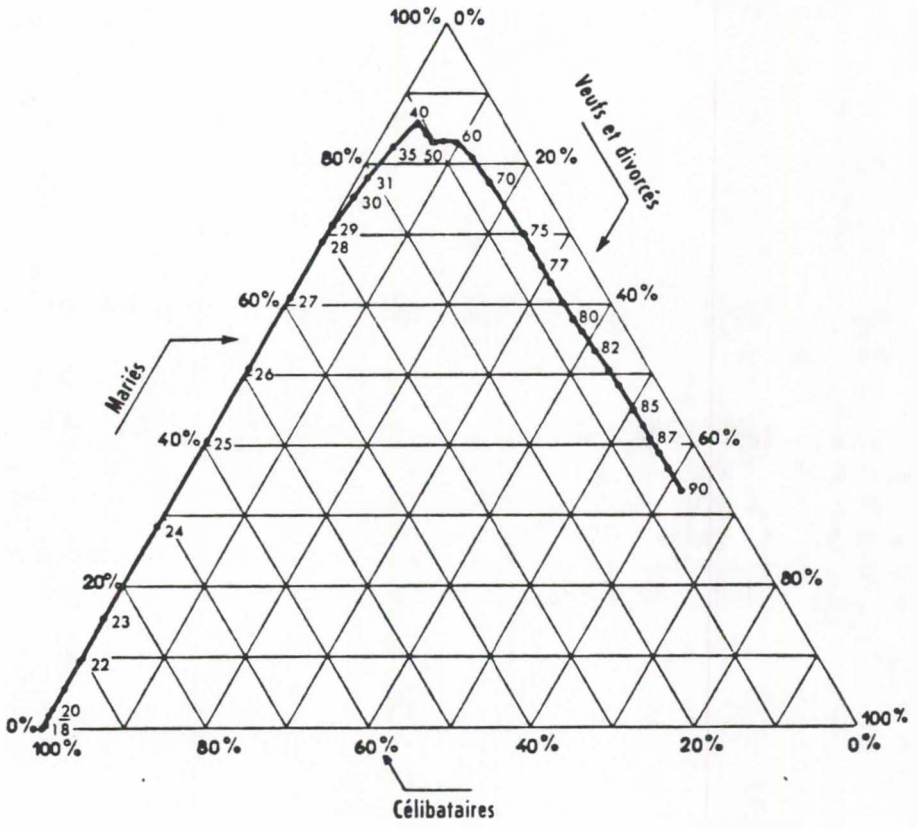


Fig.3 : Répartition de la population masculine de 1960 par âge et par état matrimonial (D'après Callot, 1984).

Un deuxième exemple (Callot, 1984) permet de présenter les possibilités de discrimination d'ensembles d'individus à l'aide du graphique triangulaire. La figure n°4 fournit la répartition des départements français suivant la structure de la population active : secteurs primaire, secondaire et tertiaire. En s'appuyant sur la figure n°4, il est possible de repérer des groupes de départements pour lesquels le même secteur d'activité prédomine.

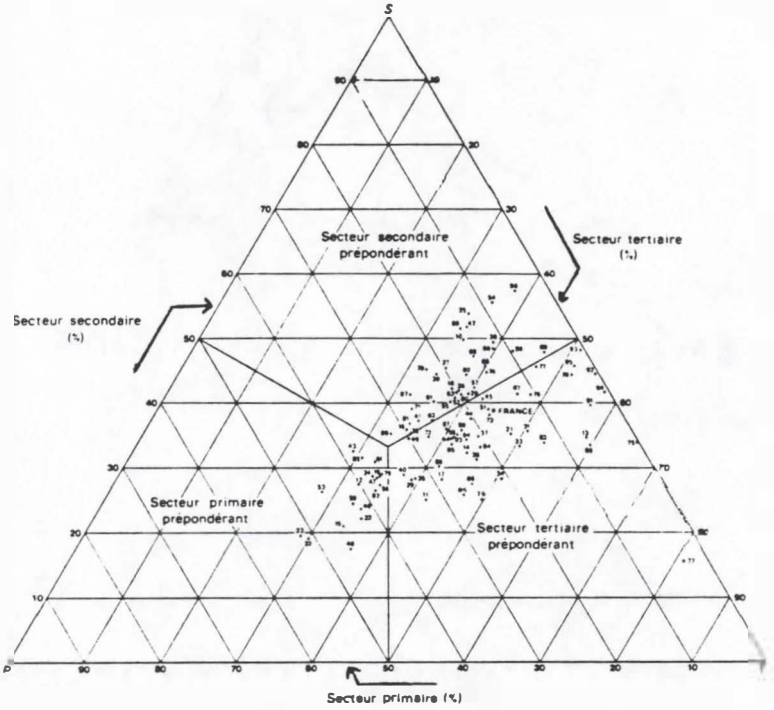


Fig.4 : Répartition des départements suivant la structure de la population active (D'après Callot, 1984).

Remarque :

Il existe une variante du graphique triangulaire qui consiste à repérer les coordonnées (α,β,γ) d'un point intérieur au triangle par les longueurs des segments découpés sur les hauteurs (fig.5). La somme de ces coordonnées est égale à chacune des hauteurs du triangle.

D'après Callot, cette représentation, qui est tout à fait identique à la précédente pour ce qui est de la position des points (seul le système de repérage des coordonnées change), est moins pratique que la première, et nécessite de repérer les coordonnées par une échelle extérieure au triangle plutôt qu'intérieure.

$$\frac{\alpha}{a} = \frac{\beta}{b} = \frac{\gamma}{c} = \cos 30^{\circ} = \frac{\sqrt{3}}{2}.$$

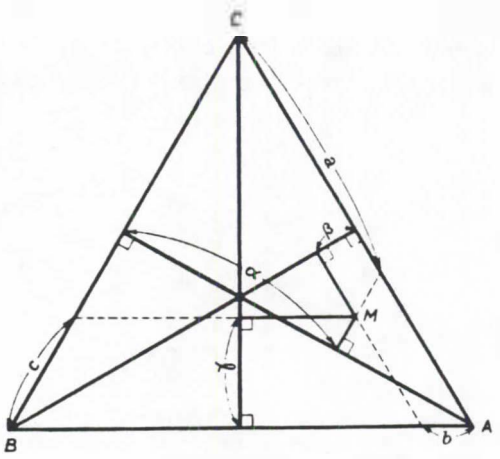


Fig.5 : Variante du graphique triangulaire (D'après Callot, 1984).

BIBLIOGRAPHIE.

- CALLOT (G.), 1984. Cours de statistique descriptive. Dunod, 488 p.

LES GRAPHES x/y ET LEURS TRAITEMENTS

Septembre 1992

Xavier PERRIER

BIOMETRIE
CIRAD - IRFA

SOMMAIRE

Document préparé principalement à partir des ouvrages :

- *Graphical methods for data analysis . 1983*
  John CHAMBERS et al.
  Duxbury Press, Boston, 395 p.

- *Exploratory data analysis . 1977*
  John TUKEY
  Addison-Wesley, 688 p.

- *La Régression . 1983*
  Richard TOMASSONE et al.
  Masson, 180 p.

INTRODUCTION

Deux variables x et y ont été observées sur une série de n individus.

La démarche instinctive est de représenter ces données sur un graphique orthogonal. C'est certainement l'un des outils les plus puissants pour visualiser totalement la dépendance entre deux variables et il peut se suffire à lui-même. A l'inverse un critère comme le coefficient de corrélation, si souvent calculé, doit obligatoirement être assorti d'un graphique pour être correctement interprété.

Deux attitudes très différentes président à la réalisation de ces graphiques :

⇒ Explorer la dépendance d'une variable "la réponse" (notée y) à une autre variable "le facteur" (notée x) ;

⇒ Explorer la relation entre deux variables sans a priori, elles ont un rôle symétrique. L'examen du graphe x/y est équivalent à celui du graphe y/x.

La lecture se fait évidemment en fonction de ces objectifs.

L'image peut parfois être plus ou moins facilement lisible en particulier lorsque le nombre d'individus est très élevé ou lorsque la dispersion gène la vision des grandes tendances. Après de brèves remarques sur la réalisation pratique de ces graphiques, seront évoquées les méthodes permettant de mieux décrire les variations de densité de l'image.

On s'intéressera ensuite aux techniques permettant d'étudier la dépendance de y en x, sans hypothèse sur la nature de cette dépendance. Si la relation est linéaire, ou peut l'être après transformation, on est à l'évidence obligé de parler de régression.

Enfin seront décrites des techniques de représentation d'une 3° variable z sur un graphe x/y.

L'esprit "analyse exploratoire des données" qui préside souvent à ces approches graphiques, impose d'évoquer, quand cela est possible, les alternatives robustes aux méthodes proposées.

Des exemples :

- FIGURE 1 : Concentration en ozone de deux villes des US. On s'intéresse simplement à la relation entre x et y, elle est manifestement croissante et la pente est légèrement supérieure à 1. L'ajout des "boîtes à pattes" suivant x et y permet de préciser la distribution de ces variables, on observe en particulier la grande dispersion des observations de Stamford.

- FIGURE 2 : Durée de vie de hamsters en fonction du pourcentage de temps passé en hibernation. L'objet est de vérifier l'effet positif de l'hibernation sur la durée de vie. Une tendance à l'augmentation mais un bruit important.

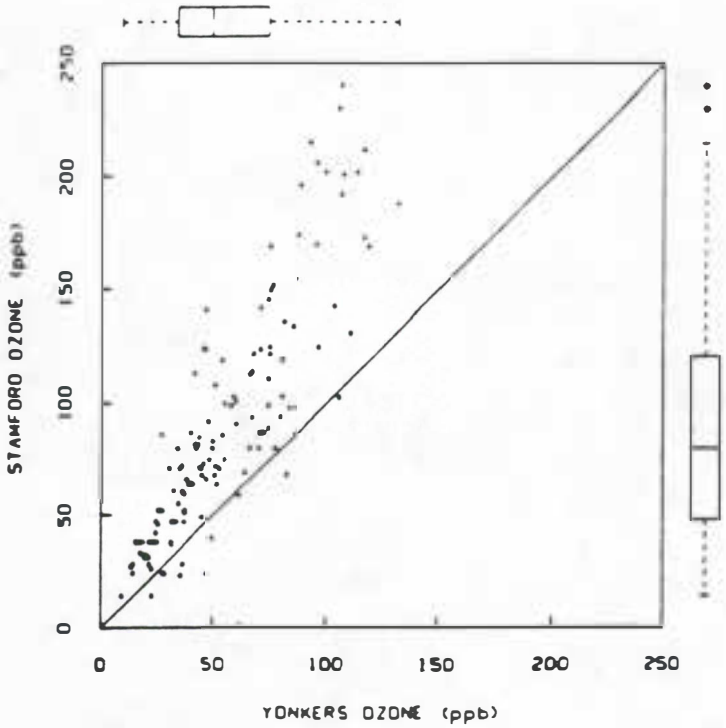


Figure 1 A scatter plot of the Stamford and Yonkers ozone data. Box plots are shown in the margins of the plot.

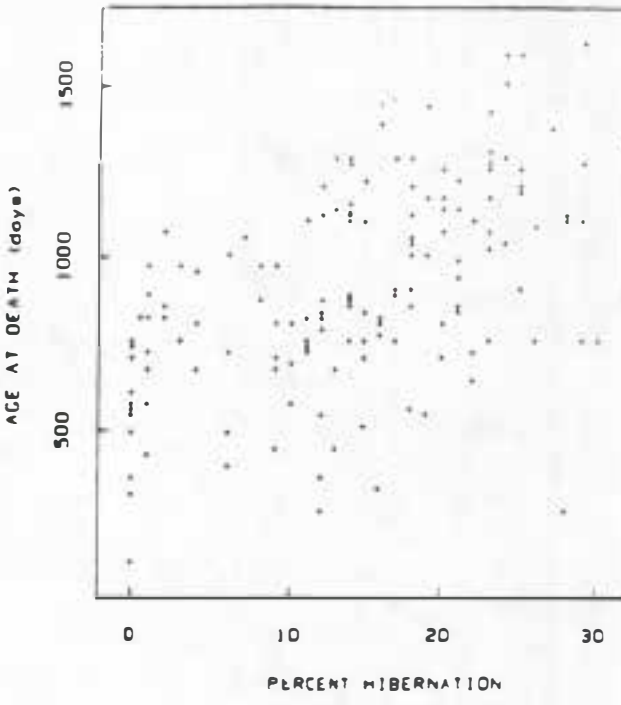


Figure 2 A scatter plot of age at death against percent o spent hibernating for 144 hamsters.

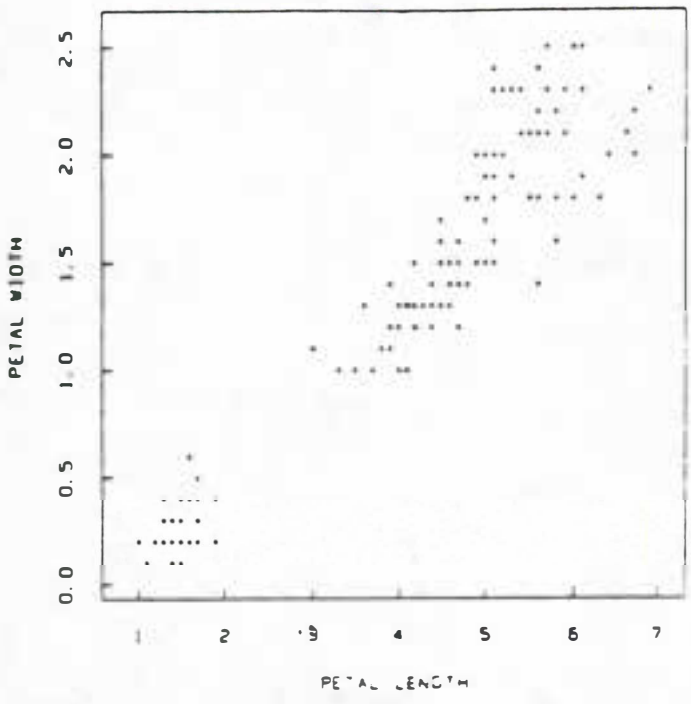


Figure 3 A scatter plot of width against length for 150 iris petals

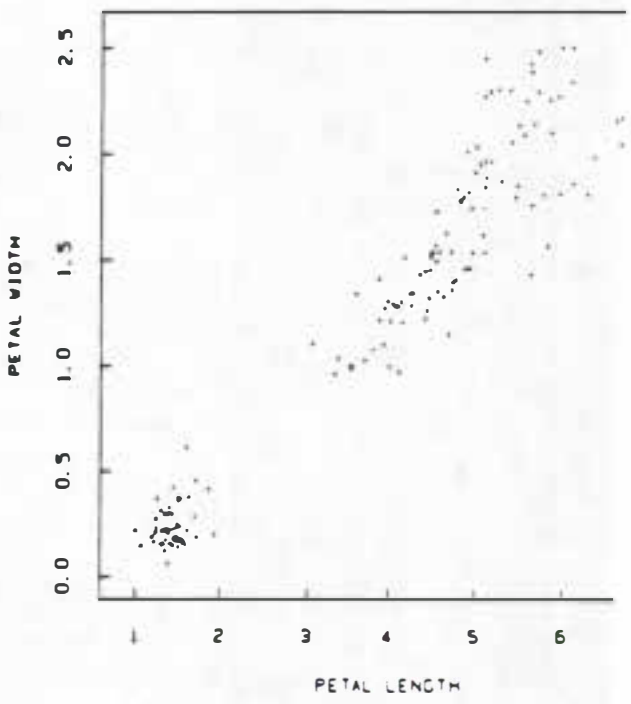


Figure 4 Petal width is plotted against petal length with jit

- FIGURE 3 : "Les Iris de Fisher" : longueurs et largeurs des pétales. On souhaite simplement décrire la relation entre ces deux variables. Deux groupes distincts apparaissent. Il y a, en réalité, 3 variétés différentes, deux sont superposées.

1 - REALISATION PRATIQUE

La réalisation manuelle d'un tel graphique ne pose pas de problème et l'on effectue, en général, d'instinct une mise en page correcte avec des échelles appropriées et des labels explicites.

La variable y est, par convention, en ordonnée. Les points ne sont généralement pas reliés sauf si l'ordre des observations a un sens, c'est le cas, en particulier, lorsque x est le temps.

Il n'est pas utile d'insister sur ces aspects qui relèvent du bon sens.

Ce bon sens n'est cependant pas aisé à mettre en oeuvre lorsque le graphique est automatisé et l'on peut évoquer quelques problèmes qui se posent :

- Il faut d'abord calculer de nouvelles échelles horizontales et verticales en fonction de l'espace physique du dessin (U_{min} U_{max} V_{min} V_{max}). On souhaite, en général, que le graphique occupe entièrement l'espace disponible. La transformation sur x est alors :

$$x_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} (U_{max} - U_{min}) + U_{min}$$

(idem sur y)

- La clarté veut que les labels sur les axes soient des nombres entiers ; une pratique commune est de choisir des valeurs qui soient des entiers consécutifs multiples d'un incrément S qui soit une fois, deux fois ou cinq fois une puissance appropriée de 10 (choisis en fonction des min et max observés).

Ex : 500, 1000, 1500 $\leftrightarrow$ $1 \times 5 \times 10^2$, $2 \times 5 \times 10^2$, $3 \times 5 \times 10^2$
(S = 1, 2, 3)

- On ne souhaite pas, en général, que les points extrêmes soient sur les axes ou sur les bordures du graphique. Une technique courante est de remplacer x_{min} et x_{max} par deux bornes fictives :

$x^*_{min} = x_{min} - 0{,}02\ (x_{max} - x_{min})$
$x^*_{max} = x_{max} + 0{,}02\ (x_{max} - x_{min})$

On "pousse" les bornes de part et d'autre d'une fraction α de l'étendue ($\alpha = 0,02$ est un valeur satisfaisante).

Dans certains cas, on peut vouloir imposer les échelles. Dans l'exemple de la figure 1 les variables x et y étant de même nature et de même unité, il est logique de forcer une même échelle dans les deux directions.

Si plusieurs graphiques doivent être comparés avec des variables de même nature, il es impératif de conserver des échelles identiques.

2 - DENSITE DES POINTS

Lorsque le nombre de points est élevé, il est difficile de saisir clairement les différent niveaux de densités de points, le cas ultime étant la superposition de nombreux individus.

2.1 - BRUITAGE ET SOLEIL

Dans l'exemple des Iris de Fisher (figure 3), l'impression générale est faussée par nombre important de points superposés, 102 points seulement sont représentés pour 150 valeu initiales.

Une solution pour faire apparaître ces points est de décaler légèrement les poin confondus par un bruitage des données (jittering). La méthode habituellement utilisée consis à transformer les coordonnées d'un point :

$$x'_i = x_i + \Theta_x U_i$$

$$y'_i = y_i + \Theta_y V_i$$

où U_i et V_i sont des valeurs aléatoires entre -1 et +1 ; Θ_x et Θ_y sont souvent pris comme u fraction (0,02 ou 0,05) des étendues $x_{max} - x_{min}$ et $y_{max} - y_{min}$.

La figure 4 reprend les données des Iris avec un bruitage Θ_x et Θ_y de 0,05 fois l'étendu Le groupe des faibles valeurs, en particulier, apparaît beaucoup plus dense.

Une technique parfois utilisée pour les données arrondies est d'utiliser pour Θ la moi de l'intervalle d'arrondi.

Une autre méthode est de donner pour les points multiples le nombre de points superposés (ce que fait le logiciel d'analyse multidimensionnelle ADDAD) ou de le symboliser par une représentation en "soleil" (sunflower).

1 point 2 points 3 points 4 points 5 points 6 points ...

Il n'est guère possible avec cette méthode de dépasser 8 points confondus.

2.2 - GRAPHES DE DENSITE LOCALE

Lorsque le nombre de points est très élevé, il peut être nécessaire d'adopter une représentation permettant une visualisation des zones de densité différente.

On peut pour cela calculer, pour chaque point x_i, y_i , une fonction de densité définie comme le rapport de la proportion de points dans un cercle de centre x_i, y_i sur la surface du cercle (figure 5 : données sur les concentrations en ozone). Le rayon du cercle est constant pour tous les points du graphique et doit être choisi en fonction des objectifs : un cercle petit donne une information très locale et la densité peut varier beaucoup entre points, même proches, un cercle grand donne des résultats beaucoup plus lissés mais est localement peu précis.

On peut donc calculer pour chaque point x_i, y_i, une densité f_i (x_i, y_i). CHAMBERS propose de définir des classes de valeurs de f_i et de réaliser un graphe x/y pour chacune des classes. La figure 6 illustre cette technique pour les concentrations en ozone. Le cercle a un rayon de 25 ppb (l'unité de concentration), les densités ont été réparties en 4 classes à partir de la médiane et des quartiles.

Une autre représentation possible sur un écran d'ordinateur serait une construction dynamique du graphique, les points apparaissant, assez lentement pour être repérables, dans un ordre de densité décroissante. L'utilisateur pourrait, à sa guise, monter ou descendre dans l'échelle de densité.

L'exemple présenté sur les concentrations en ozone est simple puisque les unités sont les mêmes sur x et y et, de plus, une échelle identique a été adoptée.

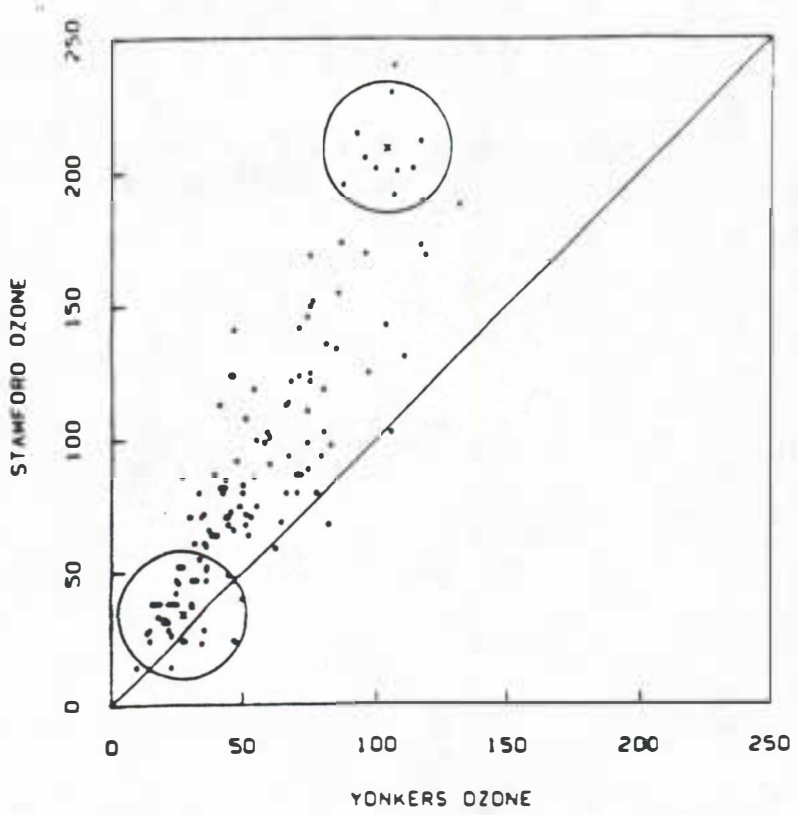


Figure 5 Stamford ozone data with two of the circles used to compute local densities.

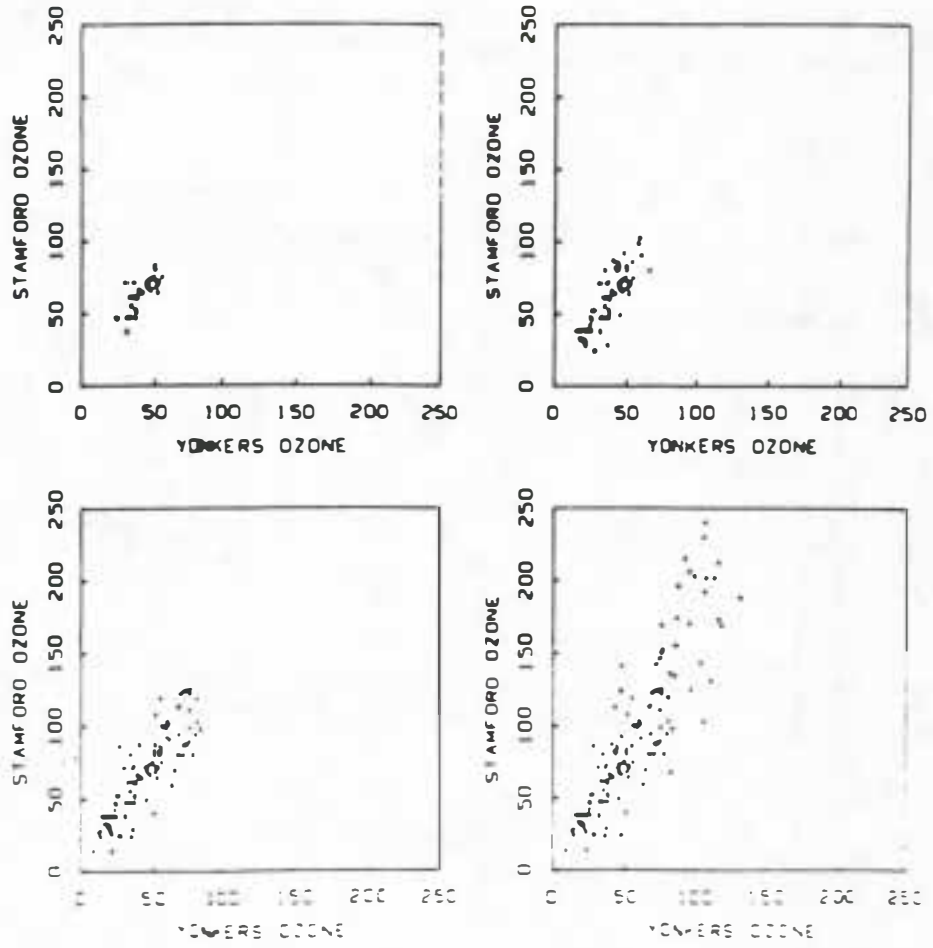


Figure 6 Four sharpened scatter plots of the ozone data

Ce n'est, en général, pas le cas. Pour tenir compte des différences d'unités, on doit au préalable standardiser les données, par exemple à partir des extrêmes et de l'étendue :

$$x^*_i = (x_i - x_{min}) / (x_{max} - x_{min})$$

$$y^*_i = (y_i - y_{min}) / (y_{max} - y_{min})$$

Pour intégrer les différences d'échelle, on choisit une unité physique de représentation (le centimètre, le pixel ou autre). Si w_x et w_y sont les étendues exprimées dans cette unité, y^*_i est corrigé du rapport w_y/w_x.

Le cercle prend alors la forme d'une ellipse allongée suivant l'un des deux axes.

3 - DEPENDANCE y/x, METHODE DES BANDES VERTICALES

Le but est d'étudier comment la distribution empirique locale de y change en fonction de x. On divise l'intervalle de variation de x en bandes avec, en général, un nombre voisin d'individus par bande (Ex. figure 7 : 5 bandes sur les données d'hibernation du hamster). Sur chaque bande, on peut estimer des paramètres de distribution : sur la figure 7, les médianes des valeurs de y dans chaque bande sont représentées par des traits horizontaux. On peut également considérer chaque bande comme une sous population et construire pour chacune une "boîte à pattes" (figure 8). Les axes verticaux de chaque boîte sont centrés sur les médianes des valeurs des x_i de la bande.

TUKEY propose de relier sur le graphique les médianes et les quantiles supérieurs et inférieurs et éventuellement de lisser ces courbes (cf chapitre 4).

Cette représentation en boîtes pour les données "hamster" permet déjà certaines conclusions : les boîtes à pattes sont assez semblables d'une bande à l'autre mis à part un glissement vers le haut. Ce glissement apparaît presque linéaire en x et concerne aussi bien la médiane que les quartiles, les dispersions sont donc comparables d'une bande à l'autre.

4 - DEPENDANCE y/x, METHODE DE LISSAGE

L'objet est de dessiner une courbe qui ajuste, au mieux, l'ensemble des points du nuage. On ne fait ici aucune hypothèse sur l'allure de cette courbe, en particulier la linéarité qui permettrait une approche de type régression.

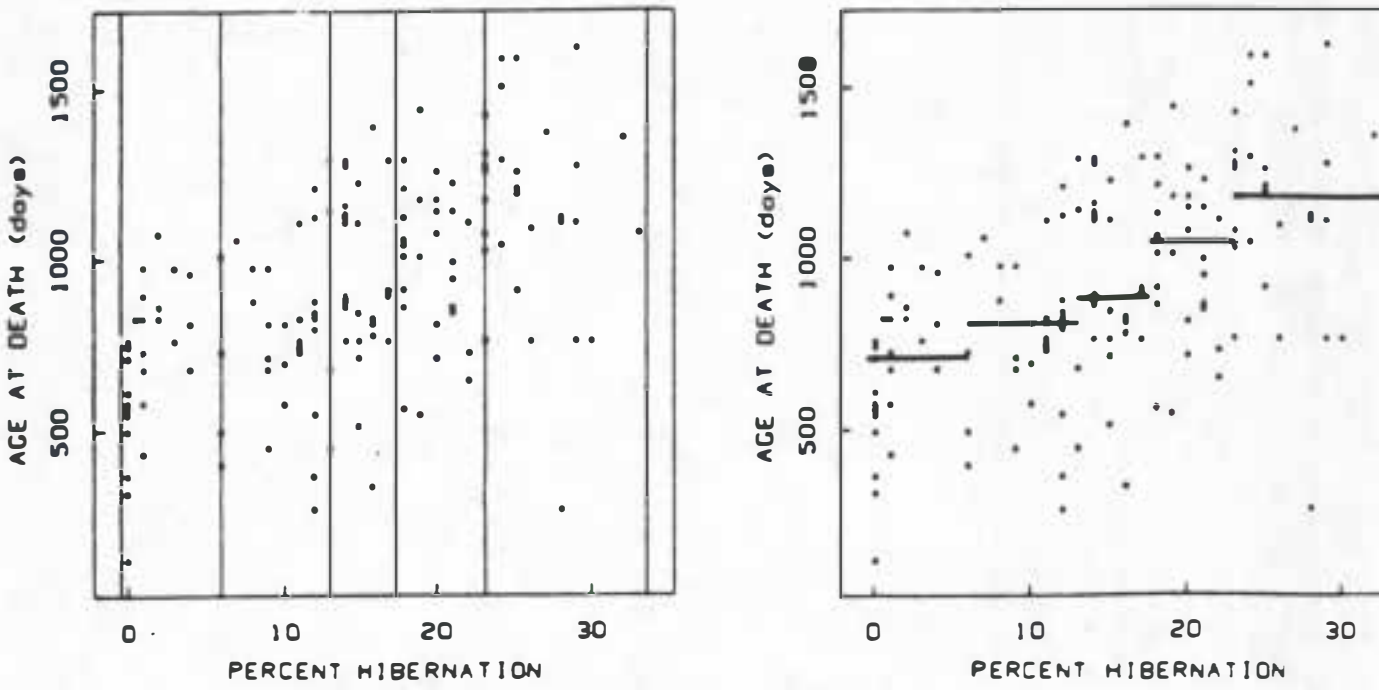


Figure 7 In the upper panel, the hamster scatter plot has been divided into strips with nearly equal numbers of points in the strips. In the lower panel, strip medians are shown by horizontal lines.

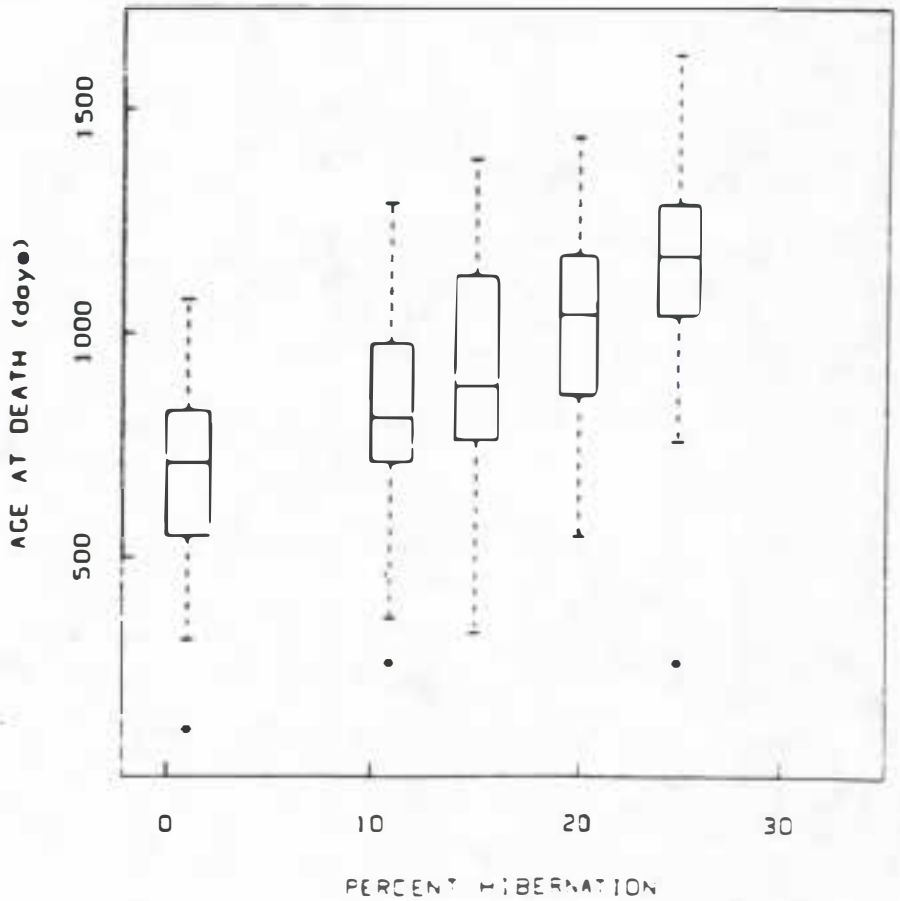


Figure 8 Strip box plots for the hamster data.

Il convient donc d'estimer, pour un couple x_i y_i, une valeur ajustée ou lissée $\hat{y}_i$ représentant la distribution de y au point $x = x_i$.

$\hat{y}_i = y_i + r_i$, r_i est alors le résidu

Si l'on disposait pour chaque x_i de plusieurs points x_i y_{i1}, x_i y_{i2},... il suffirait de choisir la moyenne ou la médiane des y_i. Ce n'est, en général, pas le cas et l'on utilise alors les points légèrement à gauche et à droite de x_i, dans un intervalle centré sur x_i : la fenêtre de lissage.

Puisque les $\hat{y}_i$ sont lissés, on relie souvent ces points par des segments de droite pour figurer la courbe lissée.

La figure 9 montre un lissage sur les données "hamsters", l'impression de croissance quasi linéaire se confirme très nettement.

4.1 - LISSAGE PAR MOYENNE OU MEDIANE MOBILE

Différentes techniques de lissage sont utilisables. La plus simple, et la plus souvent utilisée, en particulier dans les études économiques lorsque x représente le temps pour dégager le "trend", est la moyenne mobile.

On définit une fenêtre par un nombre de points, 3 par exemple. On calcule alors :

$$\hat{y}_i = (y_{i-1} + y_i + y_{i+1}) / 3$$

Cette estimation est bien sûr impossible aux extrémités où y_{i-1} ou y_{i+1} n'existent pas. On admet, en général, de perdre ces deux points. On choisit parfois de les estimer par :

$$\hat{y}_1 = (2y_1 + y_2) / 3$$

$$\hat{y}_n = (y_{n-1} + 2y_n) / 3$$

Cette méthode des moyennes déforme parfois la réalité. C'est le cas, en particulier, aux ruptures de pentes. Soient y_{i-1} et y_i, deux valeurs assez faibles et y_{i+1} une valeur très élevée ; l'estimation $\hat{y}_i$ sera tirée vers le haut par l'influence de y_{i+1}. La rupture de pente apparaîtra donc artificiellement dès x_i au lieu de x_{i+1}.

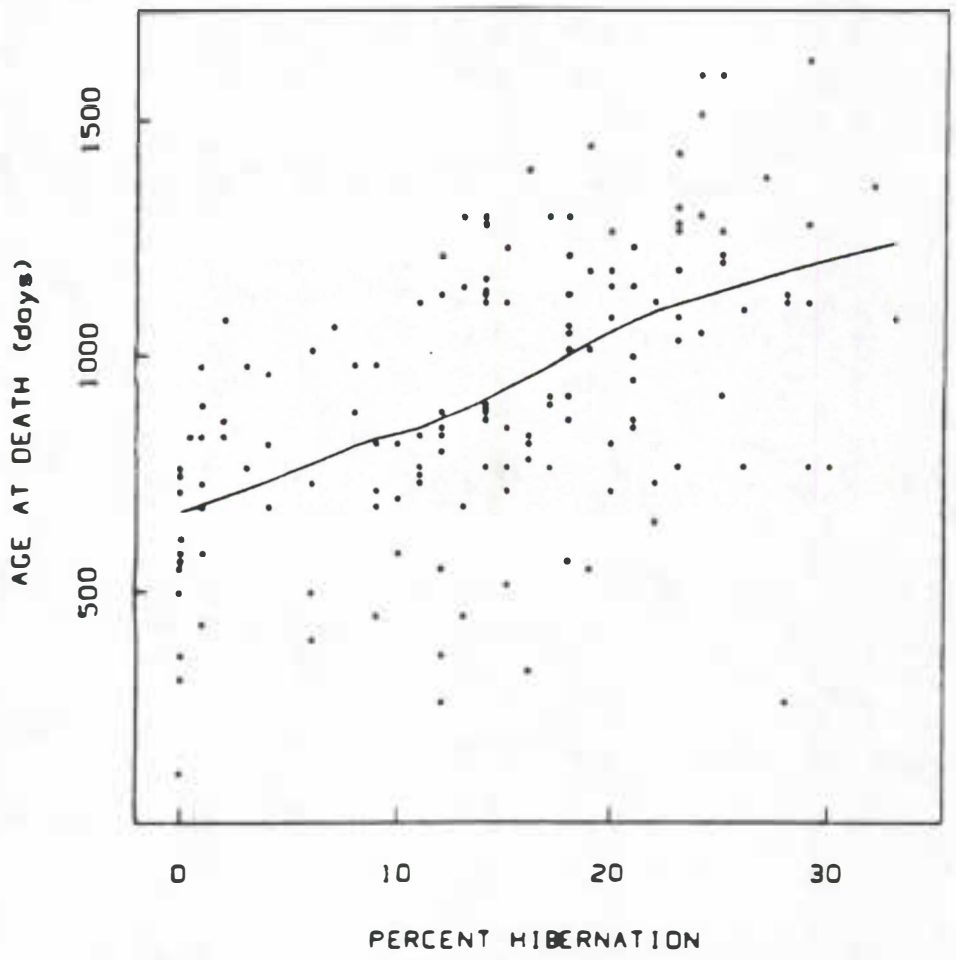


Figure 9 **The curve portrays smoothed values for the hamster data.**

On préfère souvent utiliser l'estimation des médianes mobiles :

$\hat{y}_i$ = médiane (y_{i-1}, y_i, y_{i+1})

On remarquera qu'une valeur extrême très forte ou très faible isolée (peut être une donnée anormale ?) n'apparaîtra jamais alors qu'elle aura un poids en moyenne mobile.

	y_i	8	7	9	41	13	12	8	...
(med)	$\hat{y}_i$	...	8	9	13	13	12	...	
(moy)	$\hat{y}_i$	...	8	19	21	22	11	...	

Cette fenêtre de largeur 3 est la plus souvent utilisée mais on peut, pour des raisons particulières, l'augmenter (en restant à des valeurs impaires : 5, 7, ...).

Si la courbe lissée apparaît encore trop accidentée, il est possible de répéter l'opération de lissage sur les $\hat{y}_i$.

$\hat{y}_i^* =$ médiane $(\hat{y}_{i-1}, \hat{y}_i, \hat{y}_{i+1})$

Les $\hat{y}_i^*$ peuvent à leur tour être lissées, etc...

On notera que des pics ou des creux de deux valeurs consécutives égales se conserveront quel que soit le nombre de lissage.

Y_i	...	8	9	15	15	11	7	8	...
$\hat{y}_i$		...	9	15	15	11	8	...	
$\hat{y}_i^*$			...	15	15	11	...		

On notera également que des séries strictement croissantes ou décroissantes se conserveront sans possibilité de lissage.

y_i	...	7	15	16	18	72	73	82	...
$\hat{y}_i$		...	15	16	18	72	73	...	
$\hat{y}_i^*$			...	16	18	72	...		

CHAMBERS propose certains raffinements qui permettent de résoudre ce type de problème (splitting, double lissage, etc...).

4.2 - LISSAGE PAR LA METHODE DU LOWESS

Le lissage par moyenne ou médiane mobile impose la définition de la fenêtre de lissage en nombre d'individus uniquement. Il n'a réellement de sens que si les x_i sont régulièrement espacés ce qui est souvent le cas pour les séries chronologiques mais n'est pas vérifié dans la plupart des autres cas.

L'idée de base est d'introduire la notion de distance entre individus en donnant aux points x_i de la fenêtre, un poids proportionnel à leur distance au point central, les points aux bornes n'intervenant qu'assez peu, les points du centre participant beaucoup plus activement.

La méthode souvent employée est le LOWESS pour "locally weighted regression scatter plot smoothing". Le lissage de la figure 9 a été réalisé par cette méthode.

La figure 10 détaille, sur un exemple fictif, les étapes du calcul. On s'intéresse à l'estimation de $\hat{y}_6$.

1ère étape : On définit une fenêtre centrée sur x_6 et qui comprend q individus, x_6 compris, (ici q = 10).

2ème étape : Les y_i correspondant aux points de cette fenêtre sont pondérés. La fonction de pondération prend la valeur 1 en x_6 et 0 aux bornes, elle est continue et symétrique par rapport à x_6. Les fonctions répondant à ces critères sont nombreuses. On prend habituellement une fonction tricube en posant :

$$T(U) = \begin{cases} (1 - |U|^3)^3 & \text{pour } |U| < 1 \\ 0 & \text{ailleurs} \end{cases}$$

Soit x_q le point le plus éloigné de x_i dans la fenêtre et $d_i = x_i - x_q$

Pour un point k quelconque de la fenêtre on calcule :

$$u_k = \frac{x_i - x_k}{d_i}$$

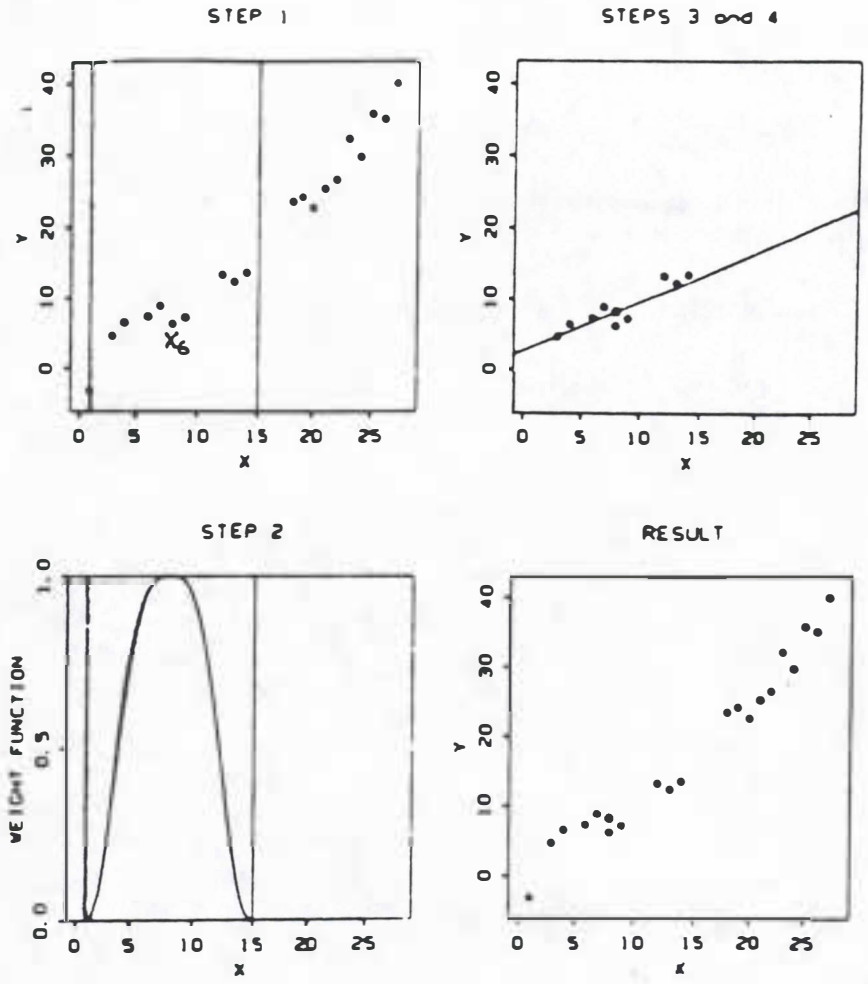


Figure 10 The four panels depict the computation of a smoothed value at x_6 using neighborhood weights.

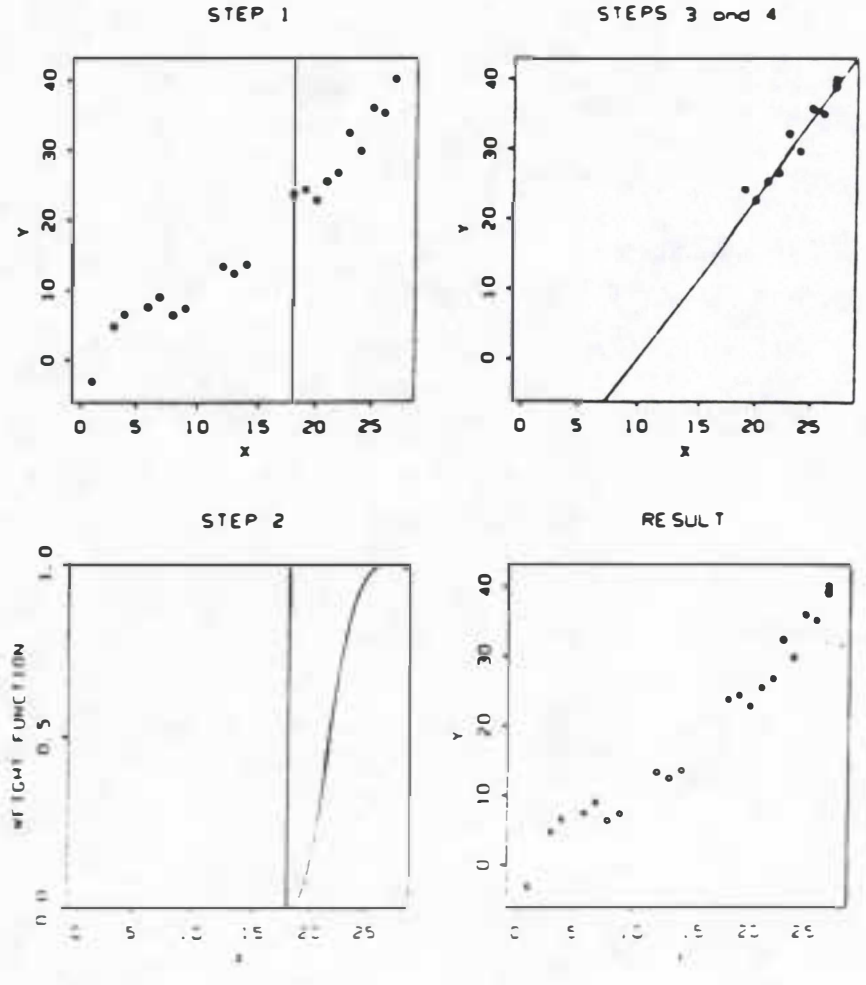


Figure 11 The four panels depict the computation of a smoothed value at [illegible]

Pour un point confondu avec x_i, $u_i = 0$.

Pour un point extrême $u_i = 1$ ou -1.

Le poids de l'individu k s'écrit alors :

$$t_i(x_k) = T\left(\frac{x_i - x_k}{d_i}\right)$$

3ème étape : On recherche la droite de régression y = a + b x sur les q points d l'intervalle par la méthode des moindres carrés pondérés, les valeurs a et b recherchée minimisent :

$$\sum_{k=1}^{n} t_i(x_k)\,(y_k - a - bx_k)^2$$

4ème étape : La valeur de lissage recherchée $\hat{y}_i$ est définie par $\hat{y}_i = a + b\,x_i$

Ces calculs répétés sur l'ensemble des points donnent la courbe lissée des $\hat{y}_i$.

Il ne se pose pas de problèmes particuliers aux extrémités. La figure 11 reprend les étap pour le dernier point x_{20}, la fenêtre est identique en nombre d'individus (10) et est donc étendu à gauche. La fonction de pondération est la même.

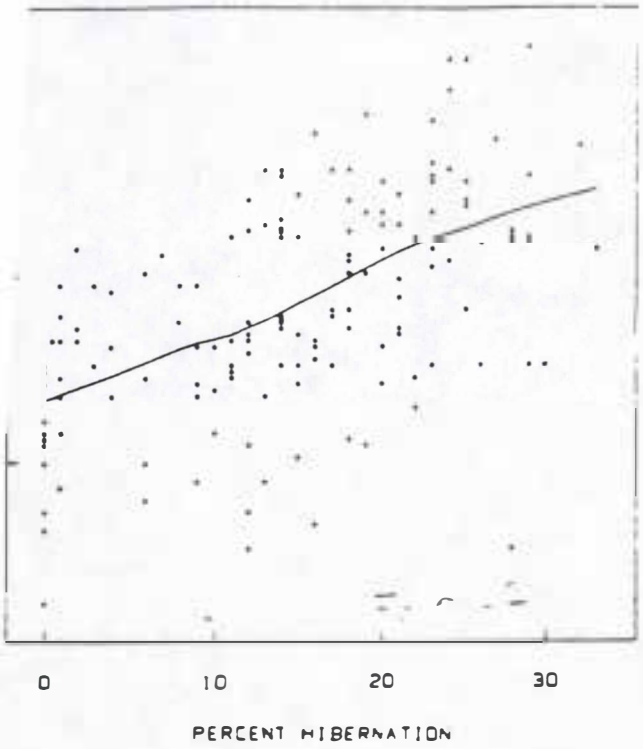
Un problème important est le choix de la taille de la fenêtre qui influe beaucoup s l'allure de la courbe. En pratique on ne choisit pas q directement mais plutôt f une fraction fix du nombre d'individus, q est alors l'arrondi supérieur de f*n.

La figure 12 donne le lowess des données Hamster avec un f de 0,67. En comparaiso la figure 13 présente les mêmes données avec un f de 0,2, l'efficacité du lissage est tr différent.

Dans une optique de graphiques dynamiques, on pourrait imaginer, sur un écr d'ordinateur, voir s'afficher successivement les courbes lissées en partant d'une faible valeur f pour arriver finalement à un f de 1.

4.3 - LOWESS ROBUSTE

La figure 13 sur les données "Hamsters" montre pour les plus fortes valeurs de x u décroissance qui contraste avec l'allure générale croissante. Elle est principalement due à point en bas à droite qui a une valeur de y très faible sans doute anormale.



12 The curve portrays smoothed values for the hamster data.

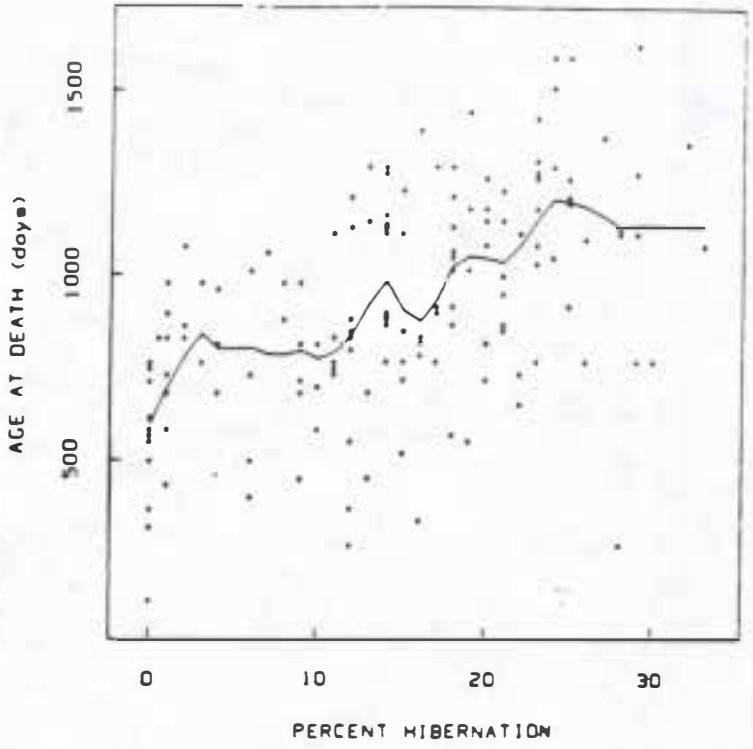
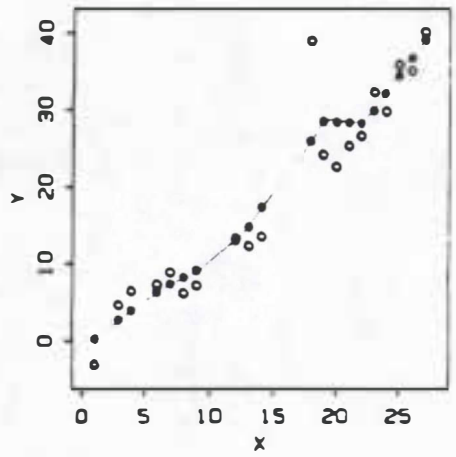
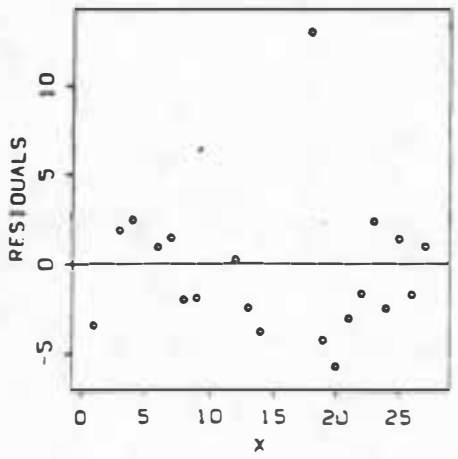


Figure 13 The curve portrays smoothed values for the hamster da with a smaller value of f





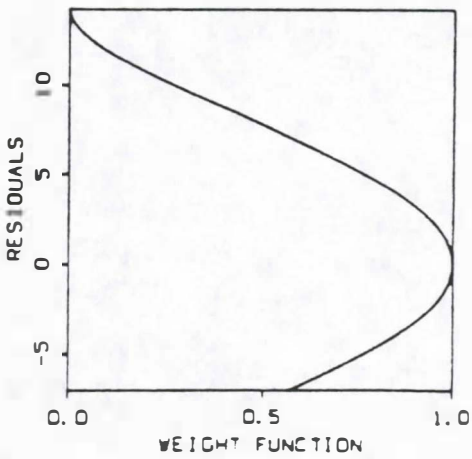


Figure 14 The three panels depict the computation of robustness weights for the made-up data with an outlier.

L'objet du LOWESS robuste est de minimiser l'influence de ces points extrêmes. Le principe de base est d'affecter aux points une nouvelle pondération qui tende à donner aux points "anormaux" un poids faible. La procédure est itérative.

1ère étape : C'est un lowess normal qui donne les estimations $\hat{y}_i$.

2ème étape : On calcule pour chaque point le résidu $r_i = \hat{y}_i - y_i$.

3ème étape (figure 14) : On affecte à chaque point une pondération qui soit proche de 1 pour les points de résidu proche de la médiane des résidus et nulle aux extrémités.

La fonction souvent utilisée est bicarrée :

$$B(U) = \begin{cases} (1 - U^2)^2 & \text{pour } |U| < 1 \\ 0 & \text{ailleurs} \end{cases}$$

Si m est la médiane des valeurs absolues des résidus, le poids de robustesse pour le point k sera :

$$b(x_k) = B\left(\frac{r_k}{6m}\right)$$

Cette valeur 6m est choisie par référence à la loi normale. Si les r_i sont normalement distribués, m étant la médiane des valeurs absolues des résidus, 50 % des points sont entre $r_i + m$ et $r_i - m$ ce qui correspond à $2\sigma/3$, σ étant l'écart type de la distribution. Si m correspond à $2\sigma/3$, 6 m correspond à 4σ. Les poids nuls n'apparaîtront donc que pour des points extrêmement éloignés.

4ème étape : On recommence l'ensemble de la procédure du lowess en ajoutant la pondération $b(x_k)$ et en cherchant a et b qui minimisent :

$$\sum_{k=1}^{n} b(x_k)\, t_i(x_k)\, (y_k - a - bx_k)^2$$

On obtient alors de nouveaux résidus, donc des pondérations de robustesse différentes. On recommence alors l'ensemble des étapes et ce jusqu'à ce que ces poids ne varient plus. En pratique, il s'avère que deux itérations sont suffisantes.

Ces méthodes de lissage et en particulier le lowess sont peu utilisées en pratique. Elles présentent cependant un intérêt particulier pour décrire l'allure d'un nuage mais aussi en préparation de traitements ultérieurs.

On notera qu'elles sont dissymétriques en x et y et que le lissage des x par rapport aux y donnerait une courbe tout à fait différente.

5 - TRANSFORMATIONS

L'interprétation visuelle d'un nuage de points est souvent facilitée si l'allure générale est de nature linéaire. Par ailleurs, les techniques de régression linéaire sont les plus efficaces et les plus accessibles. On recherche donc souvent des transformations qui permettent, quand cela est possible, de linéariser la relation entre deux variables.

La figure 15 rappelle l'allure des puissances de x et de leur logarithme les plus souvent utilisés. La forme générale produite par un lissage permet de choisir une famille de courbes de concavité satisfaisante. Des essais permettent ensuite de préciser la nature de la fonction de transformation.

On rappelera que ces transformations peuvent concerner aussi bien les x que les y. Par exemple la relation entre pression de vapeur H_20 (P) et température (T) se linéarise très bien par :

$$\text{Log}\ (P) = a + b/T$$

TUKEY propose une méthode manuelle d'estimation de la transformation : la méthode des trois points. On choisit très judicieusement (!) trois points A, B et C et l'on retient une transformation qui donne :

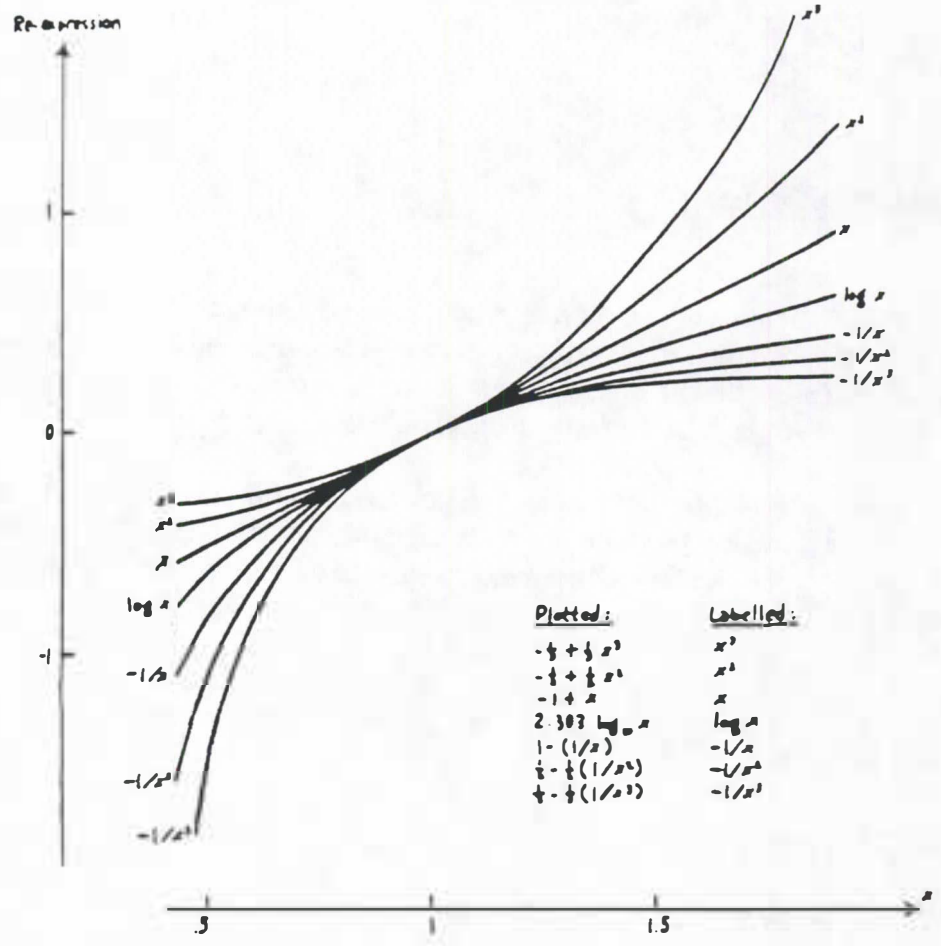
$$\text{pente}\ (A, B) \sim \text{pente}\ (B, C)$$

Cette méthode n'est valable que lorsque la transformation est de type puissance de x (ou log) ; le point B est choisi dans la zone de courbure maximale, A et C vers les extrémités en évitant les points trop extrêmes souvent plus ou moins fiables.

6 - REGRESSION

Lorsque la relation entre x et y est de type linéaire ou si le lissage nous a permis de choisir une transformation de linéarisation satisfaisante, il est classique d'estimer une droite de régression.

Il n'est pas l'objet de développer ici le thème de la régression mais différents points peuvent être considérés comme étroitement liés à notre sujet.



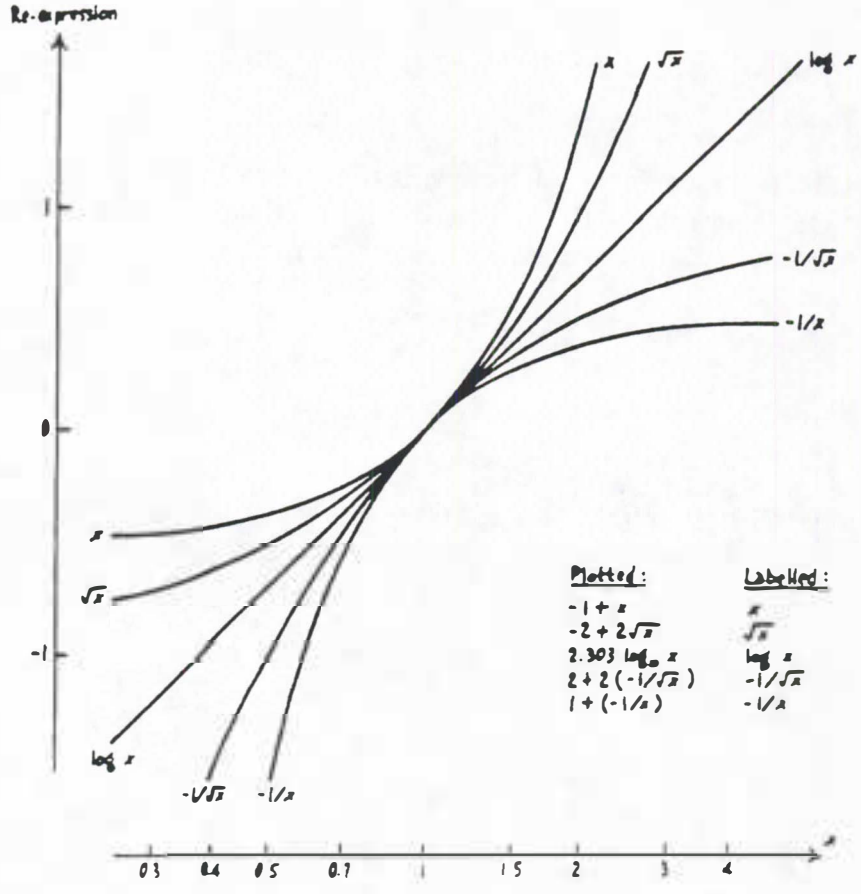


Figure 15 Principales fonctions de transformation puissance de x et de son logarithme.

6.1 - REGRESSION DE y EN x, REGRESSION ORTHOGONALE

Le principe de la régression linéaire est d'estimer les paramètres a et b d'une droite :

$$y = a + b\,x$$

telle que cette droite ajuste au mieux le nuage de points.

Ajuster au mieux signifie minimiser la distance entre un point et son estimation par la droite (le résidu r_i), c'est-à-dire minimiser :

$$\sum_{i=1}^{n} r_i^2$$

Comment estimer r_i ?

Classiquement (dans la plupart des logiciels statistiques) r_i est estimé comme la projection du point i sur la droite de régression parallèlement à l'axe des y (figure 16), c'est la régression de y en x. Cette définition n'est pas symétrique en x et y, elle privilégie x qui est fixé, c'est y qui est soumis à variation et vient se projeter sur la droite. Elle est satisfaisante si l'on s'intéresse à la dépendance de y (la réponse) par rapport à x (le facteur) (la variable dépendante par rapport à la variable indépendante dit-on en régression).

Par contre si l'on recherche simplement une description de la relation entre x et y, les deux variables jouant un rôle symétrique, cette définition n'est pas justifiée. Il est beaucoup plus raisonnable d'estimer r_i par la distance entre un point et sa projection orthogonale sur la droite de régression (figure 16), on parle alors de régression orthogonale, qui, elle, est symétrique en x et y.

On remarquera que minimiser les r_i est équivalent à conserver au mieux les distances entre deux points. Si A_b et B_b sont les projections orthogonales de A et B, $d(A_b B_b)$ est la plus proche de d(AB). Cette définition n'est autre que celle de l'axe factoriel en analyse en composantes principales. L'axe 1 d'une ACP sur deux variables x et y, est la droite de régression orthogonale.

La pente de la droite de régression de y en x s'estime par le rapport de la covariance de x et y sur la variance de x :

$$b_x = \frac{COV(XY)}{V(X)}$$

On peut définir une régression symétrique de x en y (projection parallèlement à l'axe des x) de pente :

$$b_y = \frac{COV(XY)}{V(Y)}$$

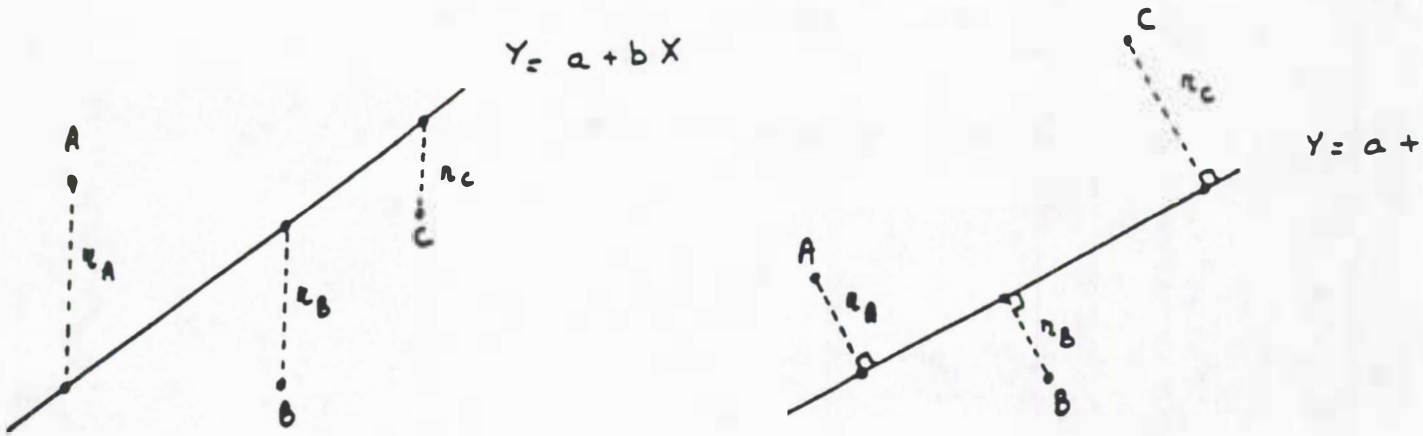


Figure 16 Régression y en x - Régression orthogonale.

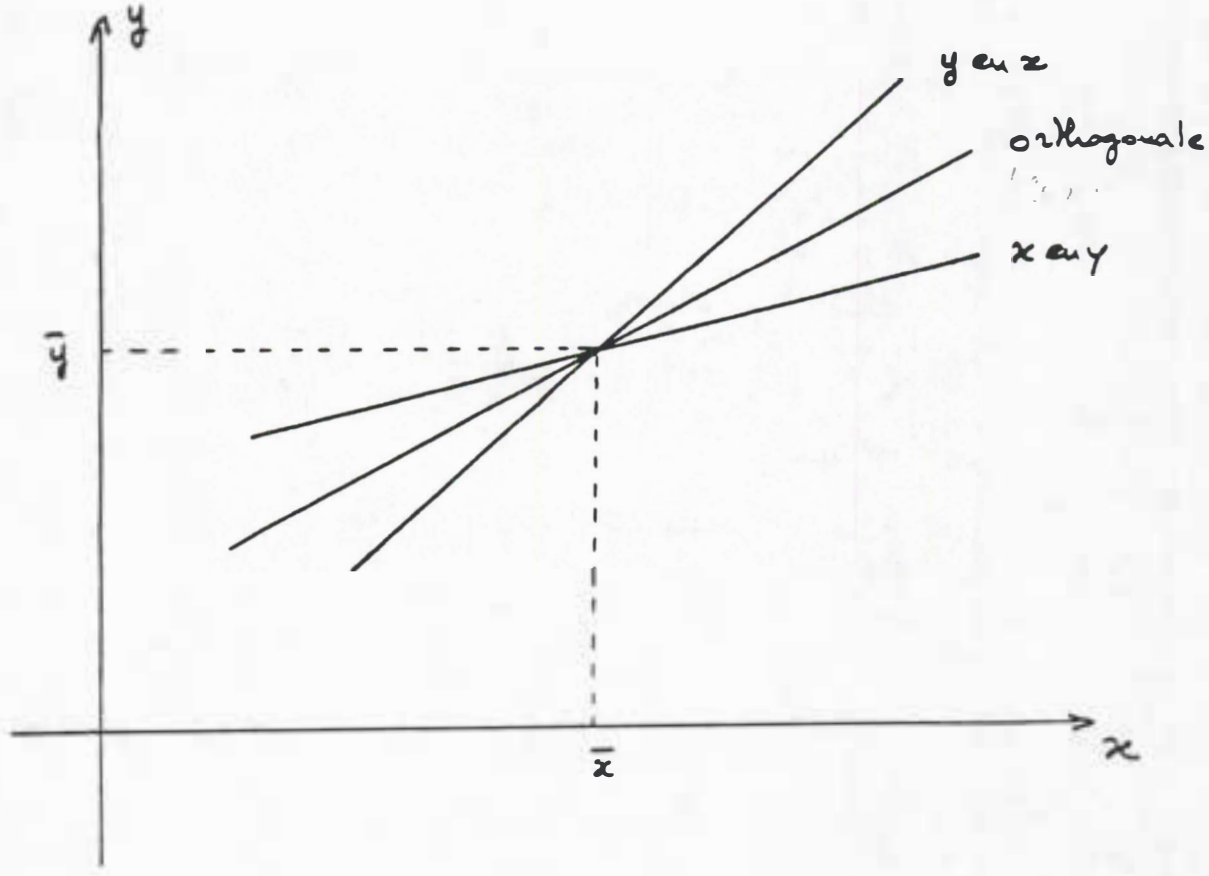


Figure 17 Les 3 droites de régression.

La pente de la droite de régression orthogonale est :

$$b_{xy} = \alpha \sqrt{\frac{V(Y)}{V(X)}}$$

α valant 1 ou -1 suivant le signe de la covariance.

On remarquera que b_{xy}, au signe près, ne dépend que des distributions marginales.

La droite de régression orthogonale est intermédiaire entre les droites de y en x et de x en y, mais elle n'en est pas la bissectrice : moyenne géométrique et non moyenne arithmétique (figure 17).

$$Y = b_x X + a \quad \textit{régression de y en x}$$

$$X = b_y Y + a' \quad \textit{régression de x en y}$$

$$\Rightarrow Y = \frac{1}{b_y} X + \frac{a'}{b_y}$$

$$Y = b_{xy} X + a'' \quad \textit{régression orthogonale}$$

$$\sqrt{b_x \cdot \frac{1}{b_y}} = \sqrt{\frac{COV(XY)}{V(X)} * \frac{V(Y)}{COV(XY)}} = \sqrt{\frac{V(Y)}{V(X)}} = |b_{xy}|$$

Ces droites ne sont confondues que si la relation est parfaitement linéaire (coefficient de corrélation = 1 ou -1) et d'équation :

$$Y = \alpha + \beta X$$

$$\Rightarrow COV(X,Y) = \frac{1}{n} \Sigma (Y-\bar{Y})(X-\bar{X})$$

$$= \frac{1}{n} \Sigma (\alpha + \beta X - \alpha - \beta \bar{X})(X-\bar{X})$$

$$= \frac{1}{n} \Sigma \beta (X-\bar{X})(X-\bar{X})$$

$$= \beta V(X)$$

de même $V(Y) = \beta^2 V(X)$

alors $b_x = \frac{1}{b_y} = b_{xy} = \beta$

Si la relation n'est pas parfaite, l'écart entre pentes peut devenir important. Il est don très souhaitable de choisir le type de régression en fonction des objectifs : dépendance ou relatio et de ne pas utiliser systématiquement, comme c'est souvent le cas, la régression de y en x.

6.2 - GRAPHIQUES DANS LA REGRESSION

De nombreuses présentations graphiques peuvent être utilisées avec profit dans le techniques de régression. Nous avons vu déjà que des méthodes de lissage pouvaient étayer de hypothèses de linéarité ou suggérer les transformations nécessaires.

Le deuxième graphique, absolument indispensable, est le graphique des résidus e fonction de la variable explicative. Il permet de déceler les écarts à la linéarité. Sur la figure 18 la régression de gauche donne des résidus élevés mais sans lien évident avec x. La régressio de droite montre des résidus nettement liés à x, une transformation s'impose. On pourrai d'ailleurs utiliser un lowess sur ces résidus pour aider au choix d'une bonne transformation.

Des techniques graphiques vues par ailleurs comme les graphes de symétrie ou d normale-quantiles permettent d'étudier la distribution des résidus et leur normalité.

Lorsque l'ajustement est réalisé, on résume souvent l'information par un coefficient d détermination (R^2), qui mesure la part de variance de y expliquée par x, et par l'écart type de résidus qui indique l'amplitude de la variation non expliquée.

Différents graphiques peuvent compléter l'interprétation.

- Graphe de la valeur ajustée $\hat{y}$ et de y. A l'évidence si l'ajustement était parfait tous l points devraient s'aligner suivant la diagonale. Sur les données d'une étude de l'élévation température de moteurs (TAR) en fonction de la vitesse du rotor (SPEED), deux régressions o été réalisées : la première sur les données brutes, la deuxième en prenant le carré de la vitess transformation suggérée par un lissage. La figure 19 présente les graphes $\hat{y}$/y pour ces deu régressions. La transformation permet un meilleur ajustement. Il faudrait s'interroger sur les points surestimés en haut et à droite et sur le point isolé à gauche ;

- Graphe des résidus ou de leur valeur absolue avec y. Les points doivent se répart sans structure, autour de la droite d'ordonnée nulle. Une forme plus ou moins courbe inciter à une transformation de y. La figure 20 montre ce graphique sur les mêmes données, on observe une tendance à l'augmentation de la dispersion des résidus pour les fortes température Ceci est plus net sur le graphique des valeurs absolues des résidus.

- Le graphique final pourra être le graphe x/y des données de départ, avec en plus droite de régression et les limites de confiance à un certain seuil.

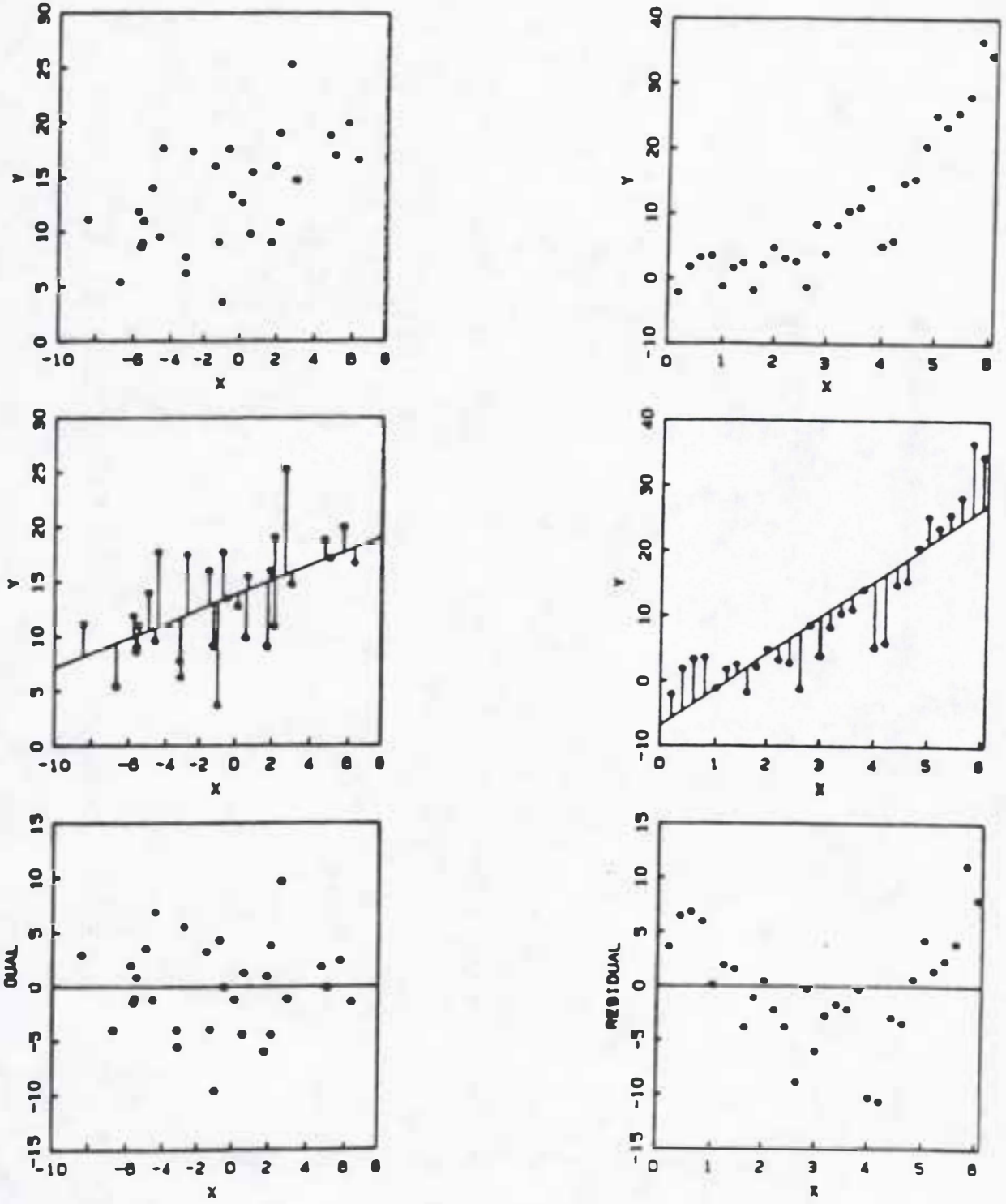
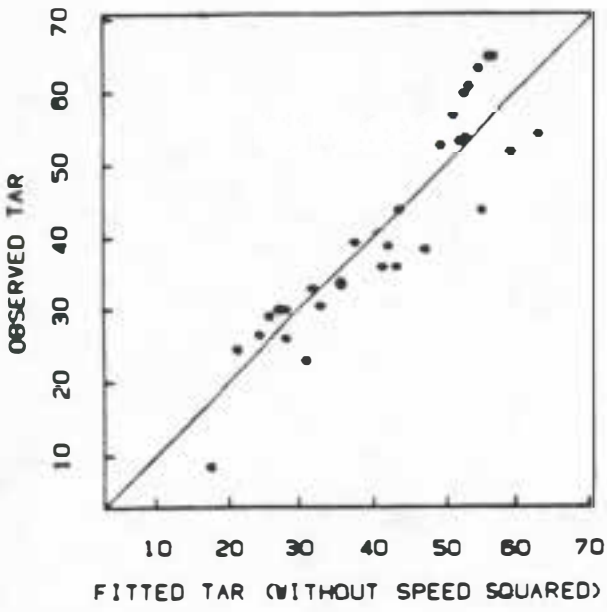


Figure 18 Examen des résidus de deux régressions.



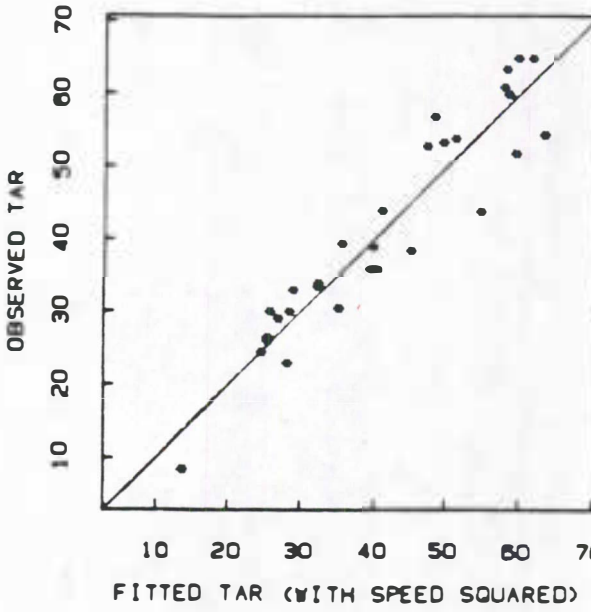
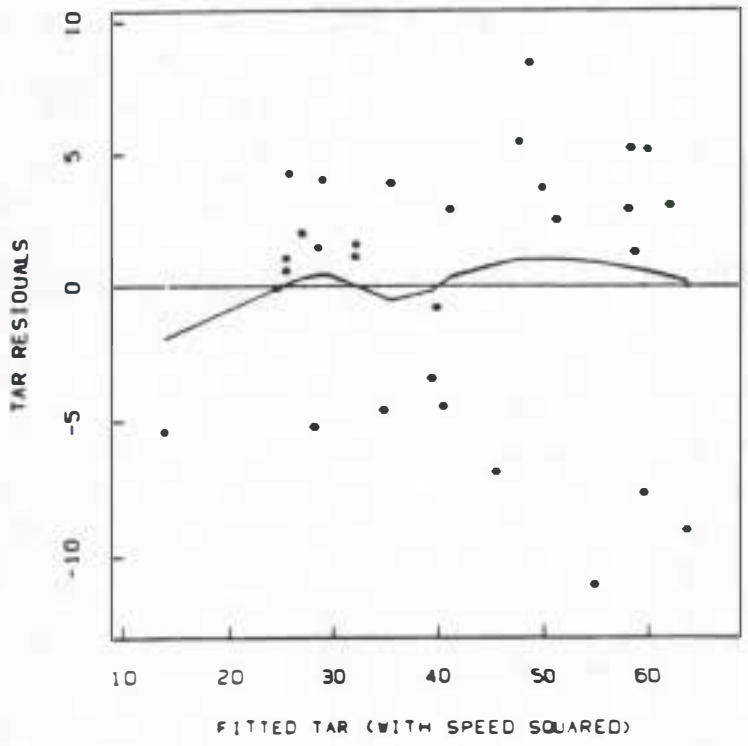


Figure 19 Graphe y observé / y estimé avec ou sans transformation x^2.



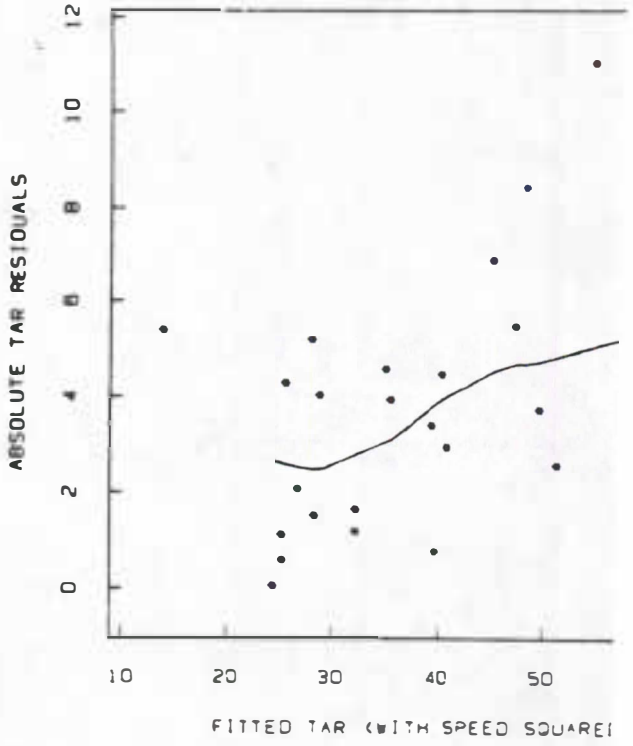


Figure 20 Graphe des résidus (et de leur valeur absolue) en fonction des y estimés.

Nous n'avons abordé que la régression linéaire simple. Des techniques graphiques appropriées peuvent également être utilisées en régression multiple pour juger des variables pertinentes et de leurs relations.

6.3 - REGRESSION ROBUSTE

Cette approche a déjà été abordée à propos du lowess. Elle consiste à pondérer les points par un poids w_i et rechercher a et b en minimisant :

$$\sum_{i=1}^{n} W_i (y_i - a - bx_i)^2$$

Dans certains cas ces pondérations peuvent être connues a priori, dans un contexte donné où l'on ne souhaite pas voir tous les points jouer le même rôle.

Souvent on n'a pas d'idée a priori sur les poids mais l'on souhaite limiter l'influence de points suspects. Comme pour le lowess robuste une première régression va permettre de calculer un résidu pour chaque point. Si m est la médiane des valeurs absolues des résidus, on définit un pseudo résidu u_i.

$$u_i = \frac{\hat{y}_i - y_i}{6m}$$

(cf chapitre lowess pour la justification)

La fonction de pondération peut être là encore la fonction bicarrée :

$$w(U) = \begin{cases} (1 - U^2)^2 & \text{si } |U| < 1 \\ 0 & \text{ailleurs} \end{cases}$$

Une nouvelle régression peut alors être calculée en estimant a', b' qui minimisent :

$$\sum_{i=1}^{n} w(x_i) (y_i - a' - b'x_i)^2$$

De nouveaux résidus seront calculés permettant la définition de nouveaux poids. Les itérations sont arrêtées lorsque la stabilité est atteinte.

Cette technique très efficace n'est malheureusement pas disponible pour l'instant dans les logiciels statistiques sur micro-ordinateur. Les puissances de calcul devraient permettre maintenant de l'implémenter sans problème.

Signalons pour finir une technique beaucoup plus simple, dans un esprit de robustesse au hypothèses de base, la régression non paramétrique. Elle consiste à calculer la pente pour tou les couples de points :

$$b_{ij} = \frac{y_i - y_j}{x_i - x_j}$$

et prendre comme pente de la droite de régression la médiane de ces valeurs.

b = médiane (b_{ij})

L'utilisation de la médiane permet évidemment de limiter l'influence des points suspects Il est possible par des méthodes non paramétriques d'estimer un intervalle de confiance pour et même comparer les pentes de deux droites.

7 - REPRESENTER UNE 3ème VARIABLE

Soit z une troisième variable à nombre de modalités limité (5 ou 6 maximum). Si variable est quantitative on peut la découper en classes à partir des quantiles par exemple. Deu techniques sont possibles : la stratification et la représentation symbolique.

7.1 - STRATIFICATION

Les k modalités de la variable z permettent de définir k sous populations, un graphiqu différent sera réalisé (en gardant les mêmes échelles) pour chaque sous population.

Cette technique a déjà été illustrée sur la figure 6 où la 3è variable z était la densi calculée pour chaque point, 4 classes avaient été définies à partir de la médiane et des quartile

7.2 - REPRESENTATION SYMBOLIQUE

Le principe simple est d'affecter un symbole à chacune des k sous populations et représenter sur le graphe x/y les points par le symbole de leur classe (figure 21).

Si la variable z est quantitative ou ordonnée, on peut jouer avec la taille des symbo pour représenter la valeur de cette variable (figure 22). Pour être lisible ce type de graphe d utiliser des symboles comme le rond ou le carré dont on perçoit bien les changements de tai

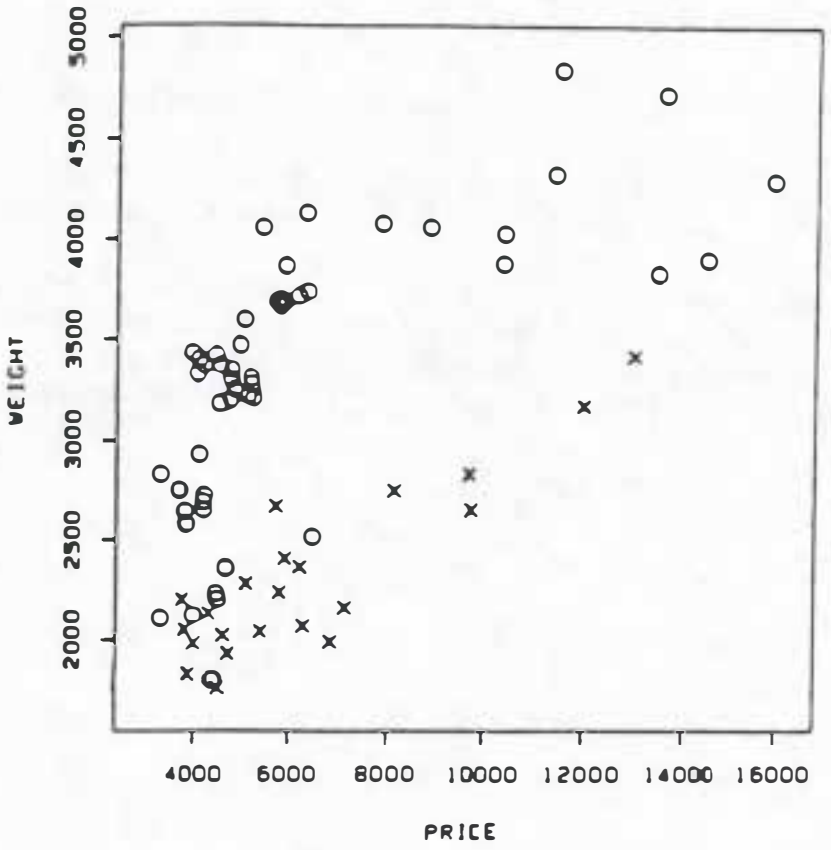


Figure 21 Symbolic scatter plot of weight against price, with U.S. cars coded as O and foreign cars as ×.

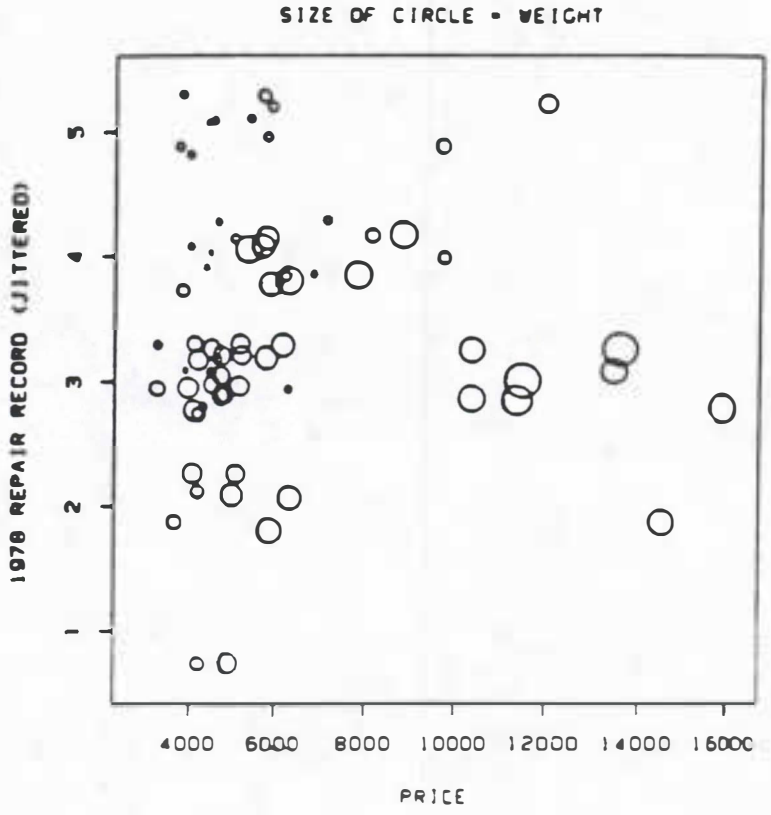


Figure 22 Symbolic scatter plot of jittered 1978 repair record against price, with automobile weight represented by the diameter of the circle.

et adopter des différences de taille nettement perceptibles entre les classes consécutives.

Dynamiquement, sur un écran, il serait possible de faire apparaître les points x y dan
l'ordre des modalités de z ou suivant la valeur de z si elle est quantitative.

Par ailleurs certains logiciels offrent une option de balayage. La variable z est symbolisé
en dehors du graphe x/y par un axe. On définit sur cet axe une fenêtre, tous les points
l'intérieur de la fenêtre s'allument (en rouge par exemple sur le graphe x/y). La souris perme
de déplacer la fenêtre ou de changer sa largeur, les représentations sur x/y sont modifiées e
conséquence.

LA REPRESENTATION GRAPHIQUE DE DONNEES MULTIVARIEES

Robert CLEROUX
Yves LEPAGE
Normand RANGER

Université de MONTREAL

RAPPORTS DE RECHERCHES

DU

DÉPARTEMENT DE MATHÉMATIQUES ET DE STATISTIQUE

Publications provisoires

LA REPRESENTATION GRAPHIQUE DE DONNEES MULTIVARIEES

par

Robert CLÉROUX, Yves LEPAGE et Normand RANGER

Janvier 1984

D.M.S. NO 84-2

C.P. 6128, SUCC. "A", MONTRÉAL, QUÉBEC H3C 3J7

Ce rapport a été publié grâce à une subvention du fonds F.C.A.C. pour l'aide et le soutien à la recherche

LA REPRESENTATION GRAPHIQUE DE DONNEES MULTIVARIEES

par Robert Cléroux(1)

Yves Lepage(2)

Normand Ranger(1)

(1) Département d'informatique et de recherche opérationnelle, Université de Montréal.

(2) Département de mathématiques et de statistique, Université de Montréal.

Introduction

La représentation graphique est utilisée depuis fort longtemps dans plusieurs disciplines. L'utilité de telles méthodes réside dans la facilité de l'oeil humain à reconnaître des formes et à associer des ressemblances ou des aberrations. Dans deux dimensions, par exemple, un nuage de points indique d'un seul coup d'oeil les relations essentielles entre les deux variables. Toutefois, dans plusieurs dimensions, même les projections dans les différents plans ne suffisent pas à simplifier la représentation graphique surtout si le nombre de dimensions est relativement grand. Il est par conséquent difficile de considérer des données dans plus de deux ou trois dimensions. Cependant, l'avènement de l'ordinateur a facilité l'exploration de données multivariées en général, et leurs représentations graphiques en particulier.

On définit l'exploration des données comme l'ensemble des techniques empiriques que l'on peut appliquer à des données avant d'effectuer des analyses statistiques proprement dites. Le but de ces techniques consiste essentiellement à découvrir les avenues intéressantes à suivre. Il s'agit presque toujours de résumer l'information par des calculs élémentaires, des tableaux, des diagrammes, des histogrammes, des graphiques. C'est une façon d'établir un contact intime avec les données brutes et ainsi, d'apprendre des choses sur ces dernières. Une exploration des données oriente souvent les analyses statistiques et consé-

quemment, les méthodes graphiques sont d'une importance primordiale.

Il existe deux grandes classes de méthodes de représentation graphique de données multivariées. D'une part, il y a les méthodes qui s'inscrivent dans une structure géométrique ou dans un modèle statistique et qui suivent une analyse des données. Toutes les méthodes dites d'analyse factorielle font partie de cette classe. D'autre part, il y a les méthodes de représentation des données brutes, celles utilisées surtout dans une démarche exploratoire. On parlera ici uniquement de ces dernières.

Toute méthode de représentation de données devrait montrer un graphique présentant certaines des qualités suivantes:

1) il doit communiquer facilement et rapidement l'information.

2) il doit aider à comprendre l'information.

3) il doit produire un impact plus grand quand les informations sont plus importantes.

4) il doit présenter un effet mnémonique en ce sens que l'information importante soit retenue par le lecteur.

5) il doit être simple, compact et attrayant.

6) il doit être clair, précis et sans distorsion.

7) il doit être facilement et rapidement compréhensible.

8) il doit être construit à l'aide de formes usuelles.

9) il doit permettre de représenter un grand nombre de dimensions à la fois.

10) il doit permettre les comparaisons et les regroupements.

Il existe un grand nombre de méthodes de représentation graphique de données brutes. Une liste succinte est la suivante: les glyphes (Anderson, 1960), les étoiles (Siegel et coll., 1971A), les polygones (Siegel et coll., 1971B), les faces (Chernoff, 1973, Chernoff et Rizvi, 1975), les courbes (Andrews, 1972), les constellations (Wakimoto et Taguri, 1978), les profils (Bertin, 1967), les triangles (Pickett et White, 1966), les étalages du dessinateur (Chambers et coll., 1983), les girouettes (Cleveland et Kleiner, 1974), les boîtes (Hartigan, 1975), les arbres (Wakimoto, 1977), les châteaux et les arbres (Kleiner et Hartigan, 1980). Le lecteur intéressé trouvera d'autres références pertinentes sur le sujet dans la bibliographie. Elles ne sont toutefois pas toutes citées dans le texte.

Dans les prochaines sections, on étudie plus en détail quatre de ces méthodes et on les applique à des données de pollu-

tion atmosphérique dans la région de Montréal,

Les données

Les données qui nous intéressent sont constituées de mesures des concentrations en parties par cent millions (ppcm) d'anhydride sulfureux, recueillies à divers postes d'échantillonnage dans la région montréalaise au cours de l'année 1975. Ces données sont échantillonnées sous l'égide des Services de Protection de l'Environnement du Québec qui en fournissent d'ailleurs des statistiques sommaires. Les résultats sont publiés par l'éditeur officiel du Québec dans les cahiers "Qualité de l'air" sous la forme de tableaux mensuels des moyennes horaires ou bihoraires des polluants. Ces tableaux comprennent aussi les moyennes quotidiennes (24 heures), les moyennes mensuelles horaires ainsi que les maximums respectifs.

Les postes choisis, au nombre de 14 sur un total possible de 17, sont distribués de façon géographiquement adéquate pour, d'une part couvrir le territoire de Montréal et pour, d'autre part, permettre l'étude de l'effet local de polluant. Ce sont:

Poste No.	Adresse	Hauteur au-dessus du niveau de la mer (mètres)	du sol	Appareil de type continu	séquentiel
1	Jardin Botanique Montréal	55	4	Titrilog	
2	Parc Jarry Montréal	55	4	Titrilog	
3	1050 St-Jean Baptiste Pointe-Aux-Trembles	-	-	Philips	
12	1125 Ontario est Montréal	23	13	Technicon	
13	1212 Drummond Montréal	35	12		Sequential
14	523 Place St-Henri Montréal	-	-		Sequential
16	7450 Champagneur Montréal	55	13		Sequential
17	10905 Berri Montreal	18	7		Sequential
20	525 9e avenue Pointe-Aux-Trembles	9	6	Beckman 906	
21	5569 Queen Mary Hampstead	55	9		Sequential
23	4330 Sherbrooke O. Westmount	55	9	Beckman 906	
24	680 Ave. Victoria Westmount	137	12		Sequential
28	rue Duncan Ville Mont-Royal	-	-	Philips	
29	Parc Pilon Montréal-Nord	-	-	Philips	

La présence d'anhydride sulfureux dans l'air est déterminée à l'aide de plusieurs méthodes: certaines stations sont équipées d'appareils automatiques qui enregistrent continuellement

les concentrations de gaz dans l'atmosphère, d'autres sont munies d'appareils séquentiels qui prélèvent des échantillons toutes les deux heures et nécessitent des analyses subséquentes en laboratoire.

Chaque poste de pollution est représenté par un vecteur de dimension 17. Les 12 premières composantes sont les moyennes mensuelles des concentrations d'anhydride sulfureux pour l'année 1975 et les 5 composantes suivantes sont déterminées par les fréquences relatives des maximums quotidiens:

composante 13: % des maximums quotidiens de 15ppcm ou plus.
composante 14: % des maximums quotidiens entre 10 et 14 ppcm.
composante 15: % des maximums quotidiens entre 5 et 9 ppcm.
composante 16: % des maximums quotidiens de 3 ou 4 ppcm.
composante 17: % des maximums quotidiens de 2 ppcm ou moins.

La représentation graphique d'un tel vecteur de dimension 17 correspond à la représentation d'un poste de pollution. La similarité entre deux représentations graphiques indique une situation de pollution semblable entre deux postes.

L'objectif de cet article étant de présenter quelques méthodes de représentation graphique plutôt que d'étudier les problèmes de la pollution atmosphérique par l'anhydride sulfureux dans la région montréalaise, seuls les postes 1, 12, 13 et 20 seront subséquemment représentés.

La méthode des étoiles

On mesure p variables sur n individus pour obtenir n vecteurs de la forme $\underline{X}'=(X_1,\ldots,X_p)$. En utilisant les coordonnées polaires, on divise le cercle en p angles égaux. Cette division définit p vecteurs du plan dont les origines se situent au centre du cercle et dont les directions sont déterminées par les angles. Pour i=1,...,p, la variable X_i est placée sur le i^e vecteur, à une distance de l'origine proportionnelle à sa grandeur. Les p variables sont ainsi placées dans le plan polaire. Ensuite, on joint les extrémités de ces vecteurs avec celles de leurs voisins immédiats pour obtenir finalement une figure polygonale qu'on appelle étoile (voir Siegel et coll., 1971A,B). On obtient donc n étoiles, une pour chaque vecteur d'observations. Il est alors possible de reconnaître des ressemblances ou de détecter des aberrations.

Dans certaines situations pratiques, les variables initiales doivent parfois être préalablement transformées afin d'assurer une certaine compatibilité entre elles. Il arrive également que certaines variantes de cette méthode soient utiles. Par exemple, on peut construire l'étoile correspondant au vecteur des moyennes, la transformer en un cercle en multipliant chaque composante des vecteurs d'observations par un facteur approprié puis, sur chaque vecteur du cercle tracer les écarts-type afin d'indiquer l'éloignement de la variable par rapport à la moyenne.

La Figure 1 indique la disposition des variables sur une étoile conceptuelle tandis que la Figure 2 présente les étoiles correspondant aux quatre postes retenus. Pour ces représentations, les composantes 13 à 17 furent multipliées par 5 afin de les rendre plus compatibles avec les autres.

On observe que le niveau de pollution est relativement faible au poste 1 (jardin botanique), fort au poste 20 (raffineries de l'est de Montréal) tandis que les structures de pollution sont semblables aux postes 12 (rue Ontario est) et 13 (rue Drummond) quoique le niveau de pollution est plus élevé au poste 13 qu'au poste 12.

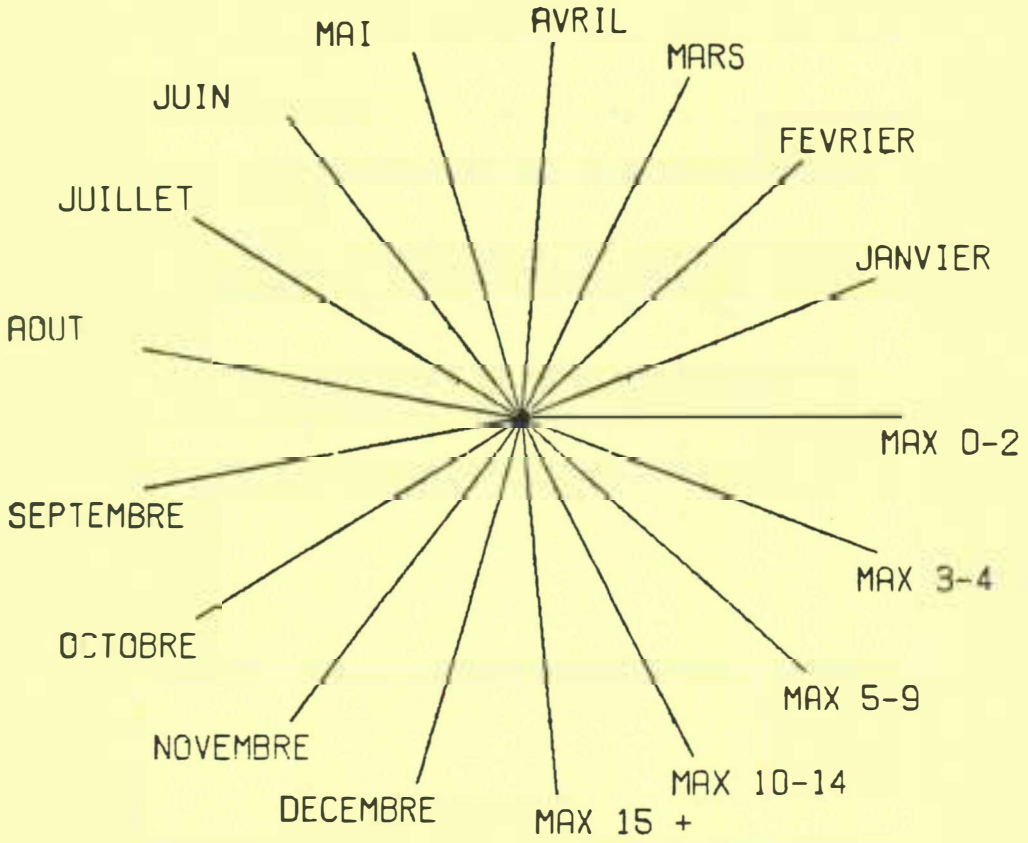


Figure 1: Disposition des variables pour une étoile conceptuelle

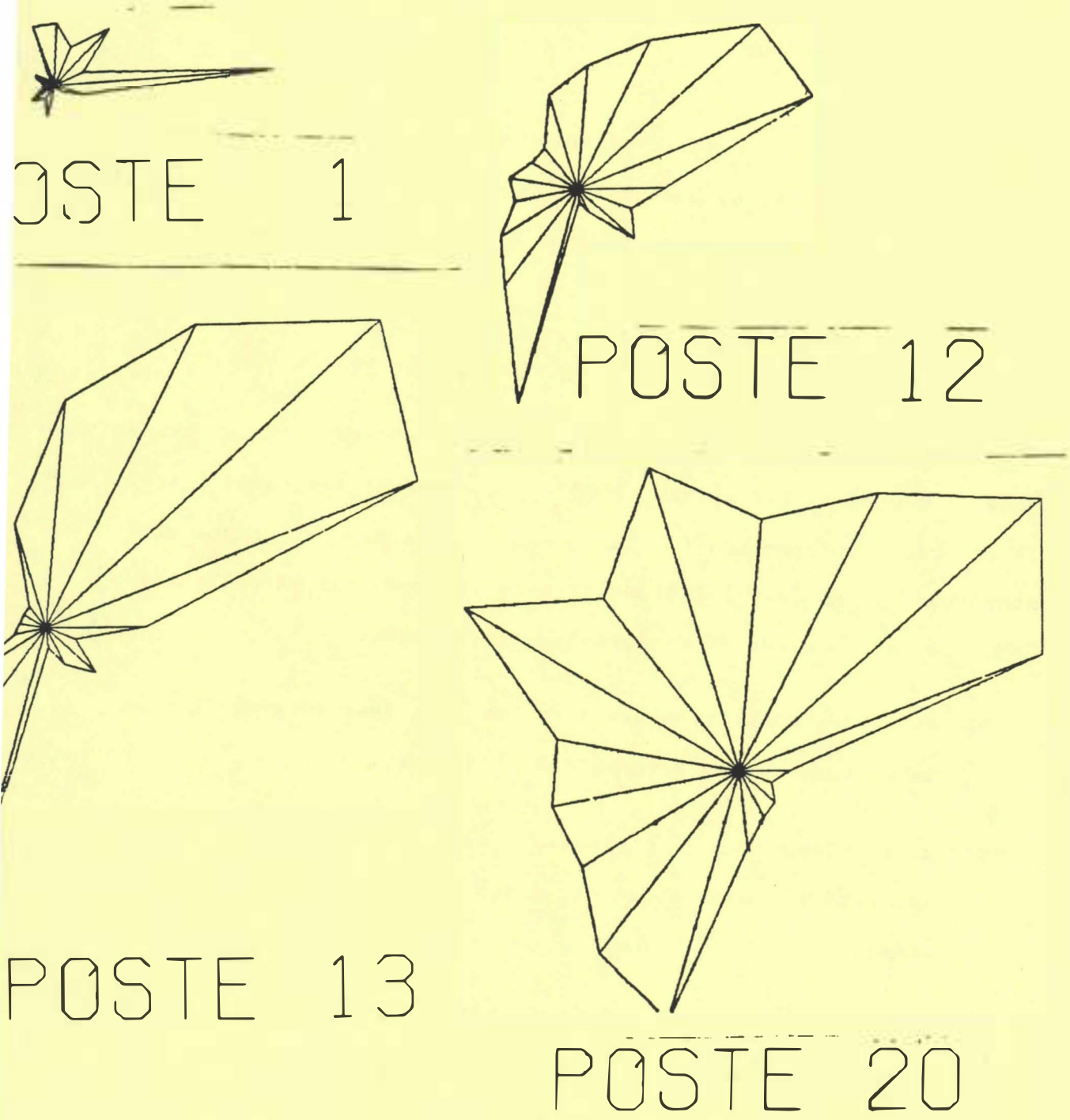


Figure 2: Représentation de certains postes de pollution par des étoiles

La méthode des courbes

On mesure n observations sur le vecteur $X'=(X_1,\ldots,X_p)$ ou, en d'autres mots, on mesure p variables sur n individus. Chacun des n vecteurs d'observations est représenté par une fonction de la forme

$$f_x(t)=\frac{X_1}{\sqrt{2}} + X_2 \sin t + X_3 \cos t + X_4 \sin 2t + X_5 \cos 2t + \ldots$$

et on en trace le graphique sur l'intervalle $-\pi<t<\pi$. On trace ainsi n courbes (voir Andrew, 1972). On pourrait également tracer le graphique correspondant à la moyenne d'un groupe et recommencer pour chaque groupe à l'étude afin de les comparer visuellement. Cette méthode possède d'intéressantes propriétés:

i) cette représentation préserve les moyennes: le graphique de la moyenne est la moyenne des graphiques.

ii) elle préserve les distances: des observations voisines (éloignées) correspondent à des graphiques voisins (éloignés).

iii) pour un t_0 donné, la valeur $f_x(t_0)$ est proportionnelle à la longueur de la projection de $(X_1,\ldots,X_p)$ sur le vecteur $(\frac{1}{\sqrt{2}}, \sin t_0, \cos t_0, \sin 2t_0, \cos 2t_0, \ldots)$. Cette propriété permet d'effectuer certains regroupements et de détecter des observations aberrantes.

iv) cette représentation préserve les variances: si les variables sont non corrélées avec variances égales à σ^2

alors

$$\text{var}\,[f_x(t)] = \sigma^2\,(1/2 + \sin^2 t + \cos^2 t + \sin^2 2t + \cos^2 2t + \ldots)$$

Si p est pair, cette variance se réduit à $1/2\ \sigma^2 p$ et si p est impair, elle se situe entre $1/2\ \sigma^2(p-1)$ et $1/2\ \sigma^2(p+1)$. Dans le premier cas, elle ne dépend pas de t et dans le second, elle dépend peu de t et la dépendance décroît quand p augmente. Donc, $\text{var}[f_x(t)]$ est approximativement constante sur tout le graphique.

v) elle permet d'effectuer des regroupements:
des fonctions voisines pour tous les t correspondent à des vecteurs d'observations voisins; des fonctions voisines en t_0 correspondent à des vecteurs d'observations voisins dans la direction du vecteur $(\frac{1}{\sqrt{2}}, \sin t_0, \cos t_0, \sin 2t_0, \cos 2t_0, \ldots)$

vi) elle préserve les relations linéaires: si le vecteur X se situe sur une droite joignant les vecteur Y et Z, alors $f_x(t)$ se situe entre les fonctions $f_y(t)$ et $f_z(t)$.

La Figure 3 montre les courbes $f_x(t)$ pour les postes retenus, en fonction du paramètre t, $-\pi < t < \pi$, Les valeurs de $f_x(t)$ sont quelque peu arbitraires mais une grande valeur de $f_x(t)$ correspond à un niveau élevé de pollution. L'axe vertical est en quelque sorte un axe de pollution.

On remarque, ici également, que le niveau de pollution est faible au poste 1 (jardin botanique) et élevé au poste 20 (raffineries de l'est de Montréal). Quant au niveau de pollution du poste 12 (rue Ontario est) et du poste 13 (rue Drummond), ils sont relativement semblables.

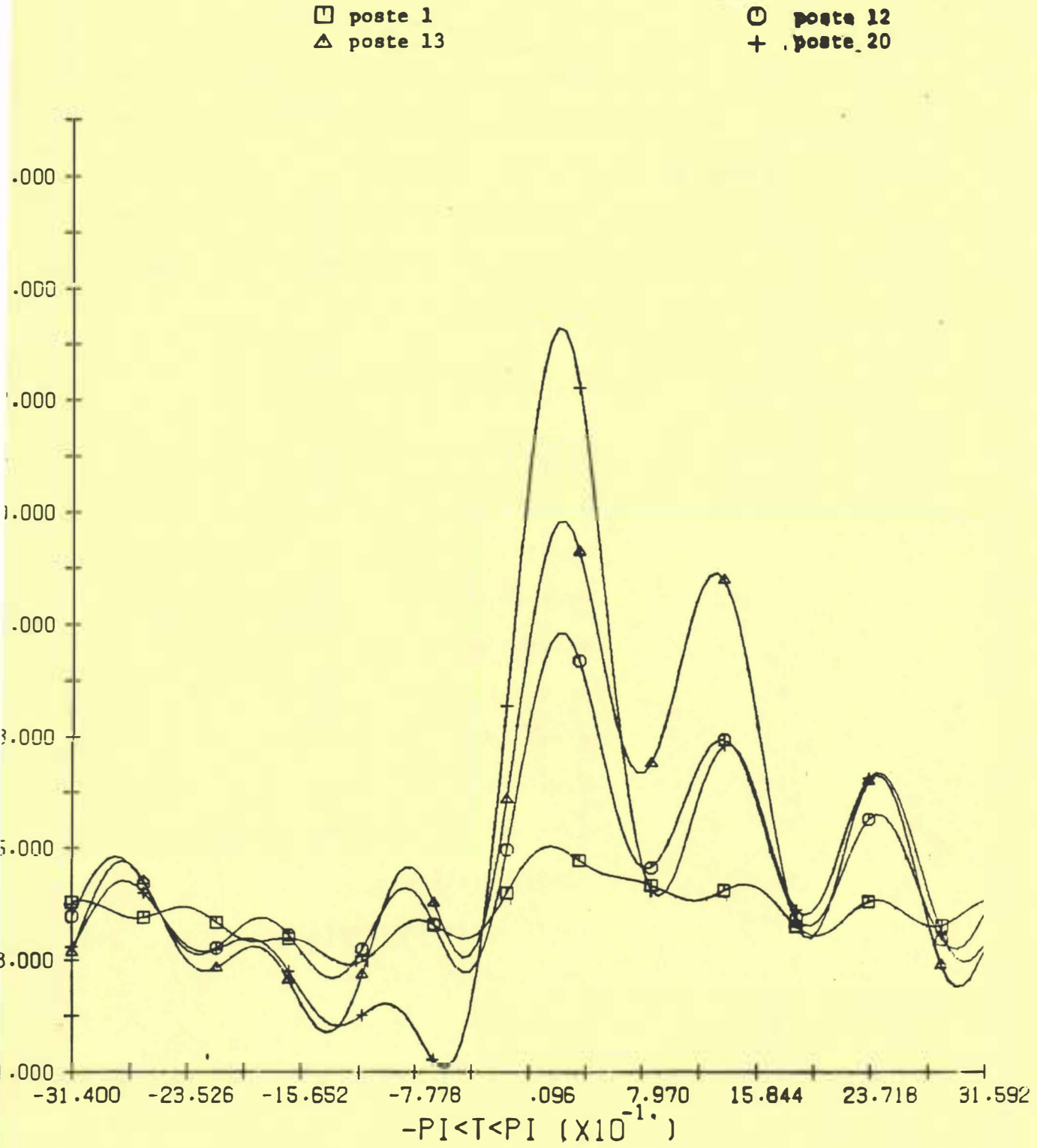


Figure 3: Courbes correspondant aux postes 1, 12, 13 et 20.

La méthode des faces

Chernoff, 1973, propose de représenter un vecteur $X'=(X_1,\ldots,X_p)$ d'observations à l'aide de faces humaines dont les caractéristiques sont déterminées par la valeur des composantes. Ainsi, à chaque composante correspond une partie de la face. Plus précisément, les paramètres utilisées pour représenter un vecteur de dimension 20 sont les suivants;

Paramètre	Partie de la face
1	largeur de la face
2	niveau de l'oreille
3	hauteur de la face
4	excentricité de la face supérieure
5	excentricité de la face inférieure
6	longueur du nez
7	niveau de la bouche
8	courbure de la bouche
9	longueur de la bouche
10	niveau des yeux
11	écart entre les yeux
12	inclinaison des yeux
13	excentricité des yeux
14	grosseur des yeux
15	position de la pupille
16	position verticale des sourcils
17	inclinaison des sourcils

18	longueur des sourcils
19	diamètre de l'oreille
20	largeur du nez

Le nez correspond à un triangle, les oreilles et les pupilles sont tracées à l'aide de cercles. Des ellipses permettent d'obtenir le contour de la face et les yeux tandis qu'un arc de cercle décrit la bouche et une droite sert à former les sourcils.

Afin d'appliquer cette méthode, il est essentiel de ramener chaque composante dans un intervalle précis pour contrôler la dimension de la face. De plus, lorsque la dimension du vecteur est inférieure à 20, une valeur préassignée est utilisée. On obtient donc n faces, une pour chaque vecteur d'observations.

Les faces de Chernoff souffrent du fait que les valeurs extrêmes de certains paramètres diminuent l'effet de d'autres paramètres (voir Chernoff et Rizvi, 1975, et Bruckner, 1978). Toutefois, conscients des limites méthodologiques, Flury et Riedwyl, 1981, présentent une nouvelle face de Chernoff permettant de représenter des observations appropriées et aussi d'augmenter la dimension du vecteur considéré.

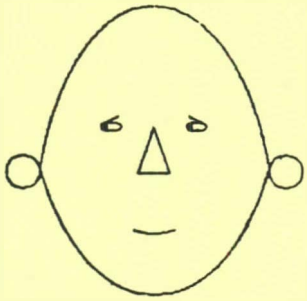
Les différents paramètres utilisés sont les suivants:

Paramètre	Partie de la face
1	taille de l'oeil
2	taille de la pupille
3	position de la pupille
4	inclinaison de l'oeil
5	position horizontale de l'oeil
6	position verticale de l'oeil
7	courbure du sourcil
8	densité du sourcil
9	position horizontale du sourcil
10	position verticale du sourcil
11	contour supérieur des cheveux
12	contour inférieur des cheveux
13	contour de la figure
14	densité des cheveux
15	inclinaison des cheveux
16	nez
17	taille de la bouche
18	courbure de la bouche

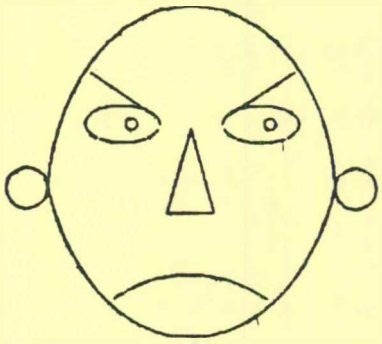
Les yeux et les pupilles correspondent à des arcs de cercles. Les cheveux, le nez, la bouche et les contours sont tracés à l'aide de courbes paramétrisées (polynômes), voir Flury, 1980. Ainsi, cette méthode permet de représenter des paires de vecteurs de dimension 18: l'asymétrie de la face obtenue reflète

le changement de chaque composante. Evidemment, elle permet aussi de représenter des vecteurs de dimension 36. Lorsque les dimensions sont inférieures à 18 ou 36 selon le cas, une valeur préassignée est utilisée. On obtient donc n nouvelles faces, une pour chaque vecteur d'observations.

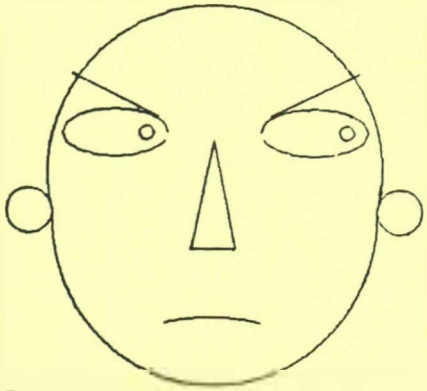
Pour les quatre postes retenus, la Figure 4 présente les faces de Chernoff tandis que dans la Figure 5, on retrouve les nouvelles faces selon Flury et Riedwyl. On observe comme par les autres méthodes, que le niveau de pollution du poste 1 (jardin botanique) est faible tandis que celui du poste 20 (raffineries de l'est de Montréal) est élevé. Les niveaux des postes 12 (rue Ontario est) et 13 (rue Drummond) sont semblables quoique le poste 13 présente un niveau de pollution légèrement plus élevé qu'au poste 12.



POSTE 1



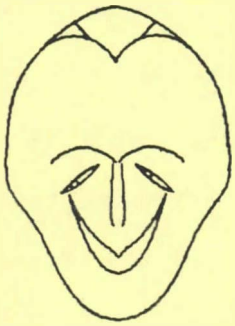
POSTE 12



POSTE 13

POSTE 20

Figure 4: Faces de Chernoff pour les postes 1, 12, 13 et 20



PO. 1

PO. 12

PO. 13

PO. 20

Figure 5: Nouvelles faces de Chernoff selon Flury et Riedwyl pour les postes 1, 12, 13 et 20

La méthode des arbres et des châteaux

Parmi les méthodes de représentation graphique de données multivariées présentées jusqu'à maintenant, l'ordre des variables du vecteur d'observations représenté joue un rôle important. En fait, chaque méthode peut offrir des aspects différents en permutant les variables étudiées. De plus, les graphiques construits jouissent d'une certaine structure de corrélation inhérente due à la construction même. Par exemple, pour les faces de Chernoff, la longueur des sourcils est fortement correlée avec la longueur des yeux; ce problème de même que celui de l'ordre des variables sont d'ailleurs traités par Chernoff et Rizvi, 1975. Afin de pallier à ces problèmes, Kleiner et Hartigan, 1981, développent une méthode utilisant un algorithme de classification hiérarchique des variables les rendant essentiellement indépendantes des effets de permutations. Cette technique d'arbres et de châteaux est connue dans la littérature anglaise sous le nom de "trees and castles".

Tout d'abord, supposons les variables $X_1,\ldots,X_p$ centrées et réduites. L'algorithme de classification hiérarchique des variables le plus fréquemment utilisé est celui à liens complets avec la distance euclidienne entre les variables (voir Hartigan, 1975). Cet algorithme de classification est préféré puisqu'il tend à diviser les variables en deux groupes de même taille. Les deux variables ayant entre elles la plus petite distance sont réunies et forment le premier groupe. La distance entre ce groupe

et chacune des autres variables est définie par le maximum de la distance entre chacune des variables du groupe et l'autre variable. Le processus est répété en réunissant la paire de groupes ou de variables avec la distance minimale, la distance étant alors définie par le maximum de la distance entre les paires. A l'étape finale, deux groupes se réunissent pour former un seul groupe contenant toutes les variables.

Pour les 17 variables observées aux 14 postes choisis, le résultat de l'algorithme de classification hiérarchique à liens complets pour les variables est présenté sous forme de dendogramme dans la Figure 6. Dans un premier temps, les variables juin et septembre sont groupées; les variables février et mars forment ensuite un groupe suivi du groupe max 15+ (composante 13) et juillet. Le processus se continue jusqu'à ce que le groupe max 0-2 (composante 17) et max 3-4 (composante 16) se joigne au groupe formé de toutes les autres variables.

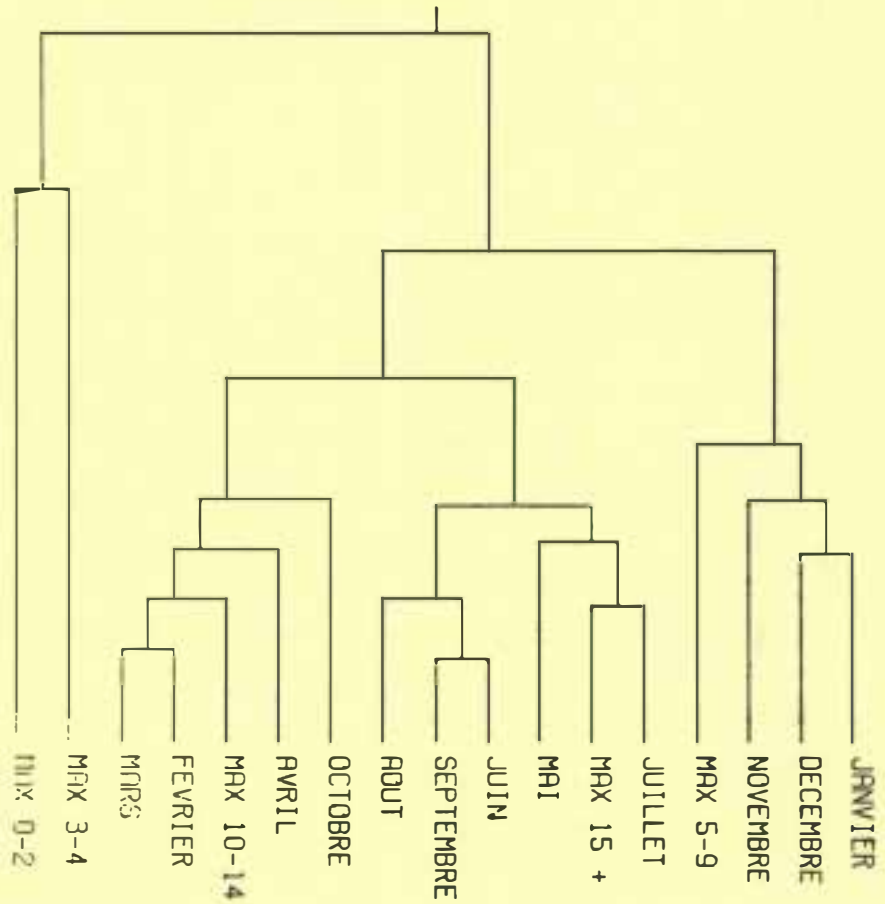


Figure 6: Dendogramme obtenu par l'algorithme de classification hiérarchique à liens complets pour les 17 variables observées aux 14 postes

La méthode des arbres représente chaque vecteur d'observations par un arbre. Tous les arbres ont la même typologie que le dendogramme des variables obtenu par l'algorithme de classification hiérarchique à liens complets et chaque feuille de l'arbre correspond à une variable.

Avant de tracer un arbre, il faut choisir la largeur et la longueur des branches de même que l'angle entre ces branches. La largeur d'une branche est proportionnelle au nombre de variables au dessus de la branche. Une variable est dite au dessus d'une branche si le trajet de la feuille représentant la variable jusqu'au pied de l'arbre emprunte cette branche. Ainsi, dans notre exemple comme chaque feuille possède une largeur de un, la base de chaque arbre est de largeur 17 tandis que la branche qui sous-tend les feuilles juillet et max 15+ est de largeur 2 augmentant à une largeur de 3 lorsque la feuille mai s'ajoute (voir Figure 7). L'angle entre deux branches à un point de division est une fonction linéaire du maximum du logarithme de la distance entre les variables au dessus de ces deux branches; le domaine de variation des angles est un paramètre à déterminer au préalable. Dans notre exemple, l'angle minimal a été fixé à 5^{o} tandis que l'angle maximal a été fixé à 95^{o}.

A chaque point de division, il faut décider de l'orientation des branches. Définissons le tronc d'un arbre par le trajet partant du pied de l'arbre et suivant, à chaque division, la branche la plus large jusqu'au point de division portant deux

branches de même largeur. Le tronc alterne de direction à chaque point de division. Au point de division ne touchant pas le tronc, la branche la plus large ira vers la droite pour les points de division à droite du tronc tandis qu'elle ira vers la gauche pour les points de division à gauche du tronc. Autrement dit, les branches les plus larges s'éloignent du tronc. Lorsque deux branches sont de la même largeur, un choix arbitraire est fait.

L'angle entre une branche et la verticale est proportionnel à la largeur de la branche sauf sur le tronc où l'angle est inversement proportionnel à la largeur de la branche. Toutefois, la somme des angles par rapport à la verticale correspond à l'angle déterminé précédemment. Ainsi, le tronc aura tendance à suivre la verticale. Enfin, la longueur d'une branche est proportionnelle à la valeur moyenne des variables au dessus de cette branche.

Dans la Figure 7, les arbres correspondant aux 4 postes de pollution retenus sont présentés. On remarque ici qu'en général, le niveau de pollution est faible au poste 1 (jardin botanique) et élevé au poste 20 (raffineries de l'est de Montréal). Bien que moins élevé que pour le poste 20, le niveau de pollution du poste 13 (rue Drummond) est plus élevé que celui du poste 12 (rue Ontario est).

Cette représentation graphique à l'aide d'arbres est très utile lorsque les variables étudiées se divisent en groupes

de variables fortement correlées. Il est de plus souhaitable que les mesures des variables soient comparables. Toutefois, cette méthode permet difficilement de comparer les variables d'un même arbre même si elles sont voisines. Kleiner et Hartigan, 1981, suggère de représenter chaque point par un château. Cette méthode est un mélange de la méthode des arbres et de la méthode des profils (voir Bertin, 1967) et elle permet de comparer facilement les variables d'un même arbre. Elle permet aussi de comparer les variables pour chaque observation plus facilement que la méthode des profils.

Supposons que les variables prennent des valeurs positives comparables. Un groupement hiérarchique à liens complets utilisant la distance euclidienne permet d'obtenir un arbre des variables. Cet arbre sert de base à la construction des châteaux. On pose la largeur d'une branche proportionnelle au nombre de branches au dessus d'elle, l'angle entre toutes branches est fixé à zéro, l'ordre des variables est celui obtenu par l'algorithme de groupement à liens complets.

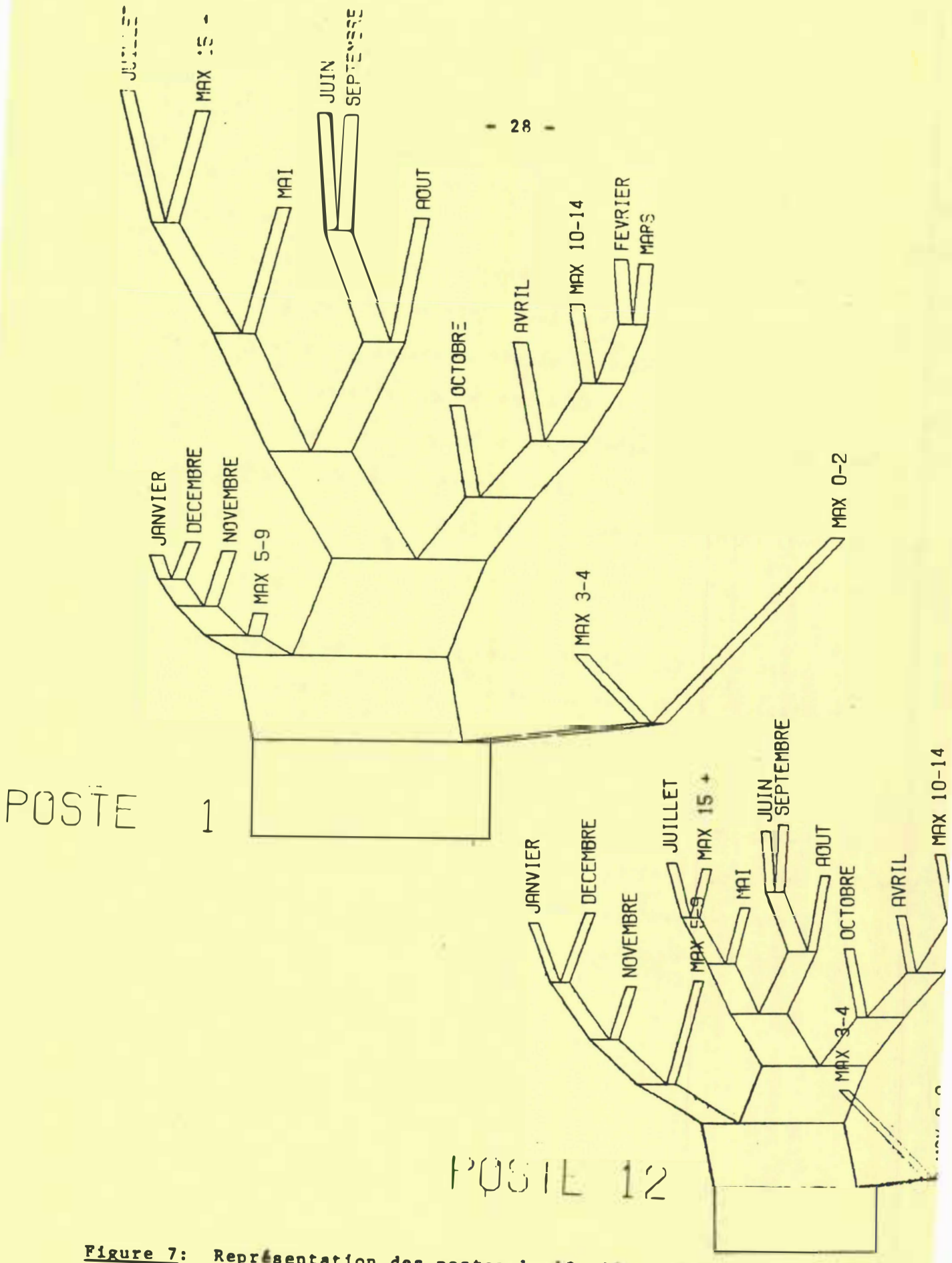
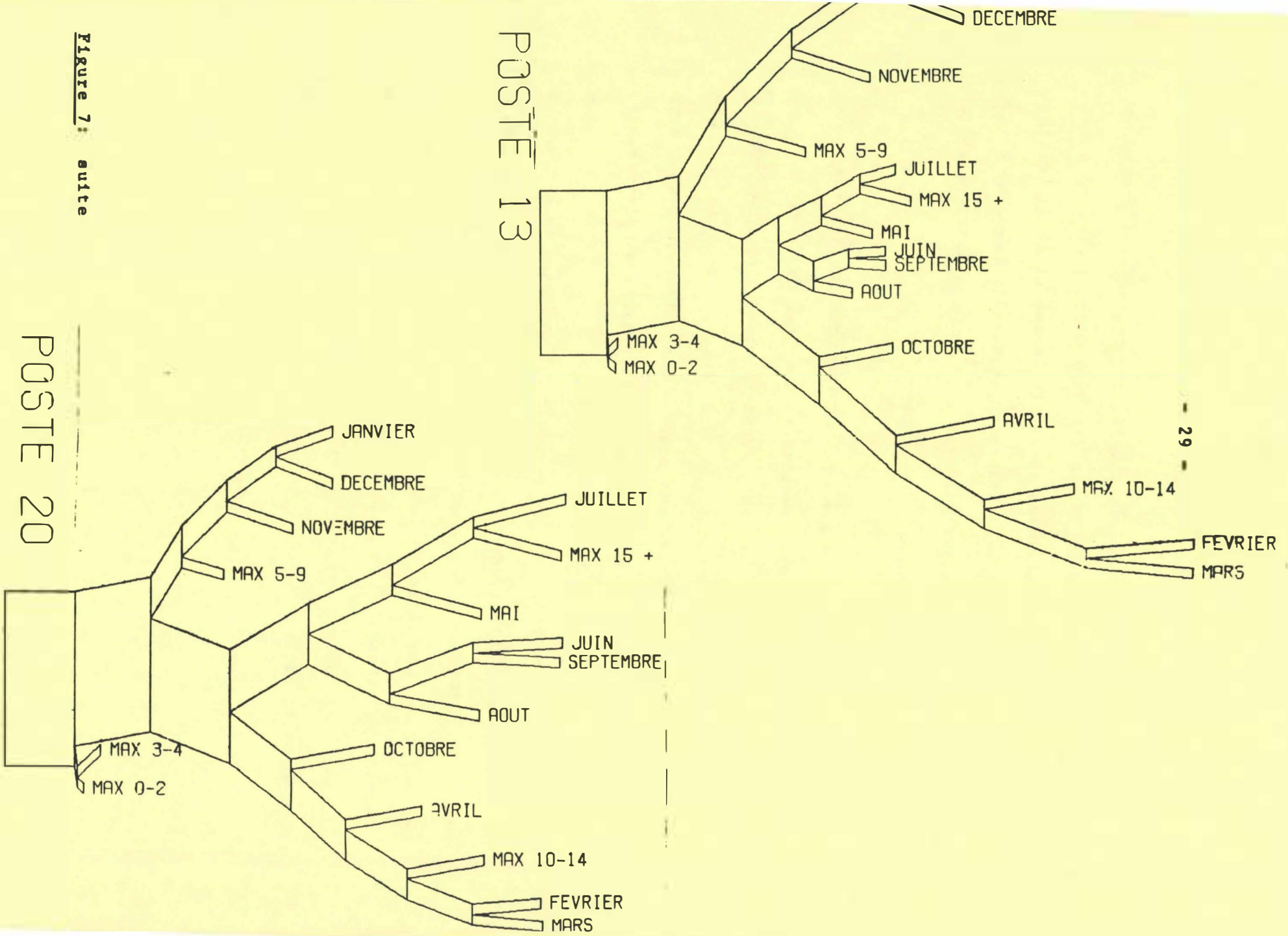


Figure 7: Représentation des postes 1, 12, 13 et 20 par des arbres

Figure 7 : suite



Ainsi, chaque vecteur est représenté par un château: l'extrémité supérieure d'une branche est à une distance v de la base où v est la valeur minimale de toutes les variables au dessus de la branche moins qd où q est le nombre de branches joignant la branche à la variable possédant la valeur minimale et d est une valeur à choisir.

La distance de la base à l'extrémité d'une branche contenant une seule variable correspond à la valeur de la variable et par conséquent, cette branche donne la même information que la méthode des profils. La position des autres branches reflète l'information de la méthode des arbres. La valeur de d doit être strictement positive afin que la structure d'arbre apparaisse.

Dans la Figure 8, on retrouve les châteaux pour les 4 postes de pollution retenus. On observe encore une fois que le poste 20 (raffineries de l'est de Montréal) possède le niveau de pollution le plus élevé tandis qu'il est le plus faible au poste 1 (jardin botanique). Le niveau de pollution du poste 13 (rue Drummond) est plus élevé que le niveau de pollution du poste 12 (rue Ontatio est).

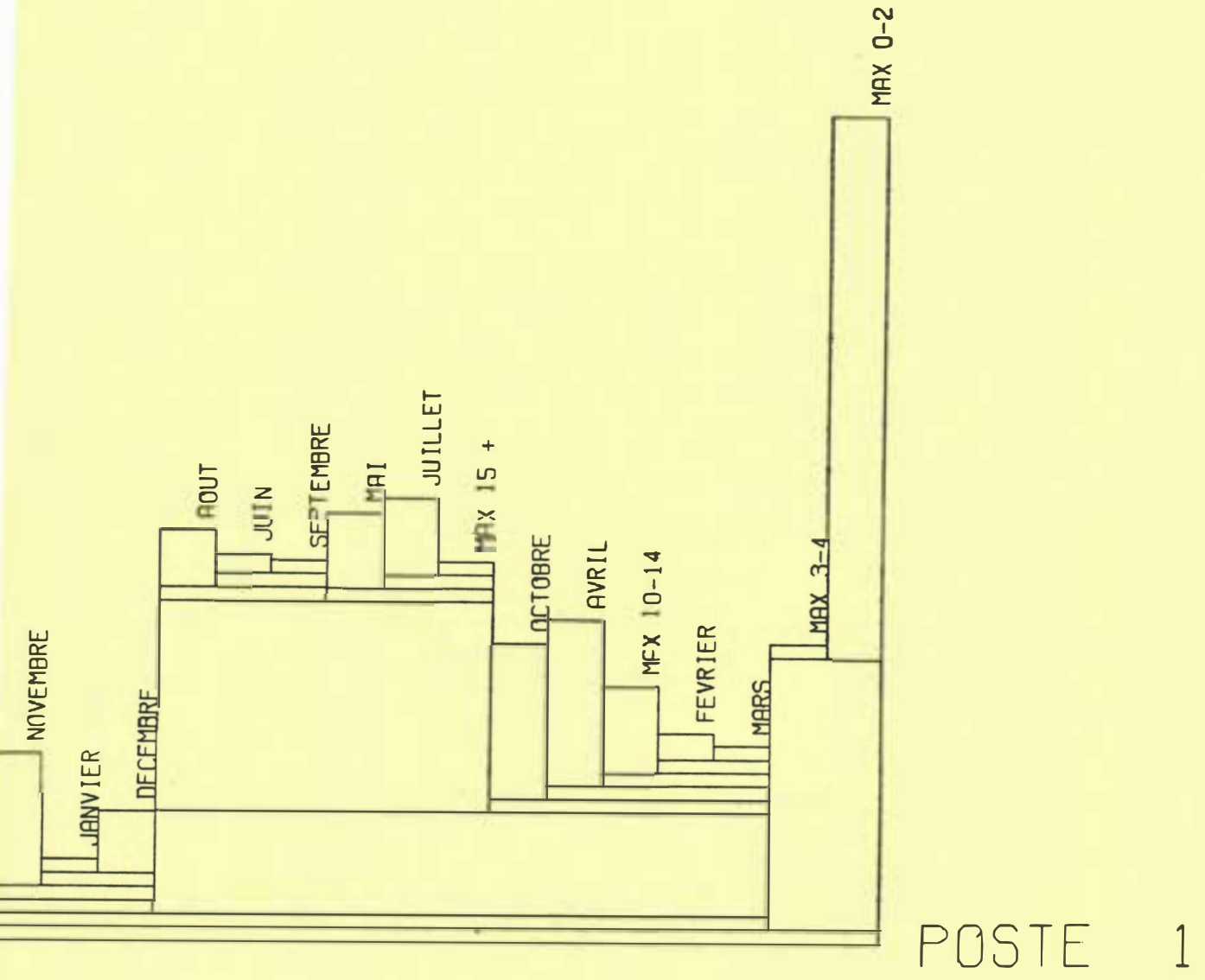
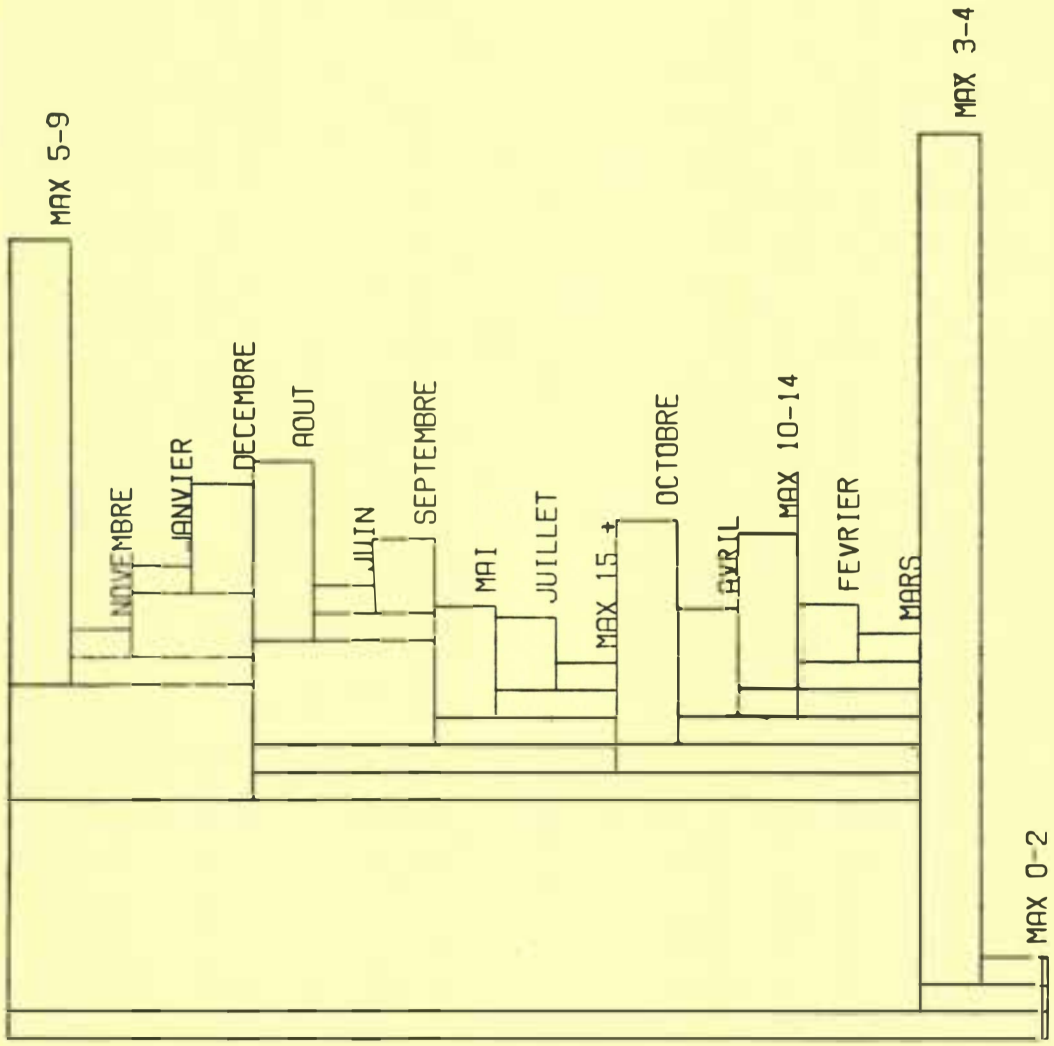
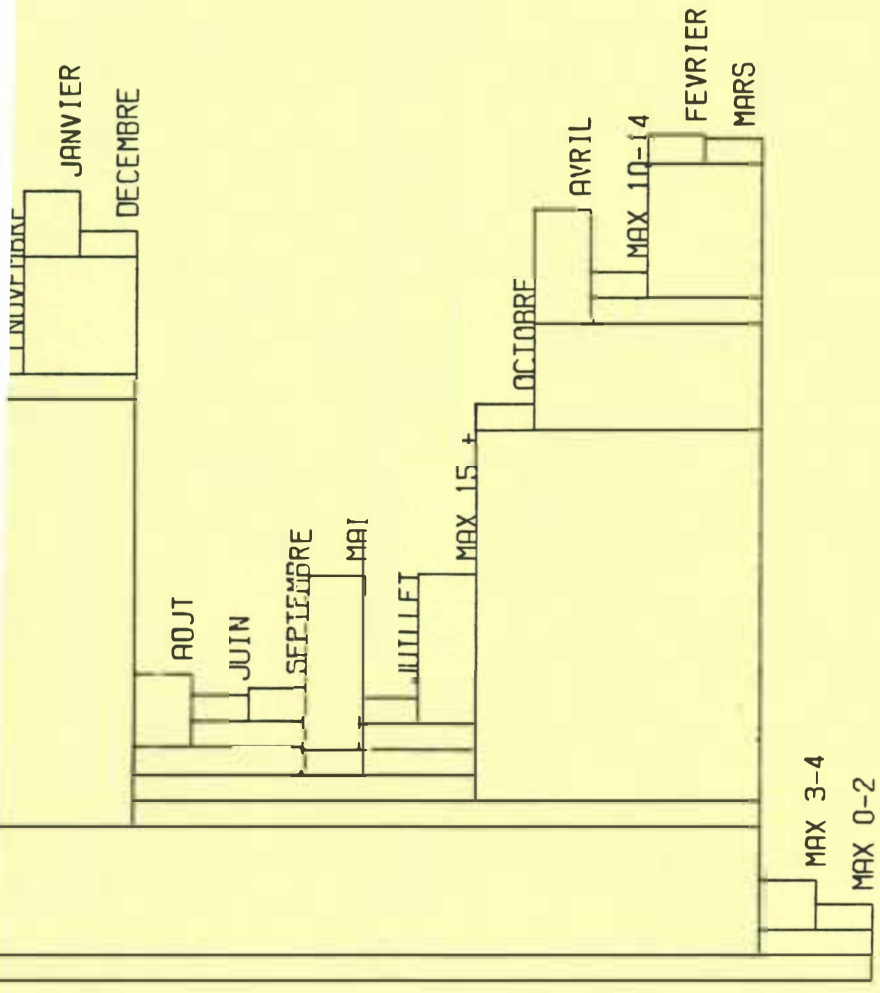


Figure 8: Représentation des postes 1, 12, 13 et 20 par des châteaux



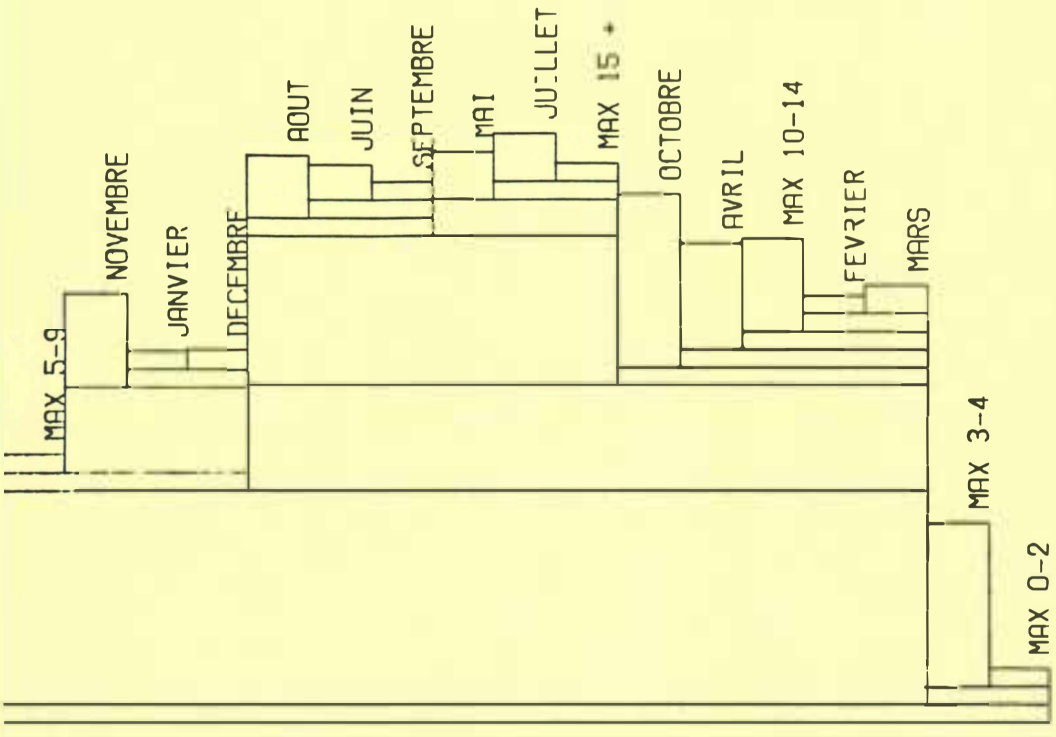
POSTE 1

Figure 8: suite



POSTE 13

Figure 8: suite



POSTE 2

Figure 8: suite

Conclusion

Dans cet article, on a esquissé quelques méthodes de représentation graphique de données multivariées. De grands pas ont été franchis depuis les travaux de Playfair, 1801, un des pères de la représentation graphique statistique. Il reste toutefois que les problèmes ne sont pas encore complètement résolus. Il faut continuer à développer, expérimenter et utiliser d'autres méthodes en se rappelant qu'un graphique est un moyen simple de communiquer les informations contenues dans un ensemble de données.

Bibliographie

Anderson, E. (1960): A Semi-Graphical Method for the Analysis of Complex Problems, Technometrics 2, 287-292.

Andrews, D.F. (1972): Plots of High-Dimensional Data, Biometrics, 28, 125-136.

Banfield, C.F. et Gower, J.C. (1980): A Note on the Graphical Representation of Multivariate Binary Data, Applied Stat. 29, 238-245.

Beniger, J.R. et Robyn, D.L. (1978): Quantitative Graphics in Statistics: A Brief History, Amer. Statistician 32, 1-11.

Bertin, T. (1967): Semiologie graphique, Gauthier-Villars, Paris.

Boivin, C. (1982): Quelques méthodes de représentation graphique des données multidimensionnelles, Directeur de la recherche, Ministère du revenu, Gouvernement du Québec.

Bruckner, L.A. (1978): On Chernoff Faces, in Graphical Representation of Multivariate Data, ed. P.C.C. Wang, New York: Academic Press.

Caporal, P.M. et Hahn, G.H. (1979): Computer Offerings for Statistical Graphics, An Overview, Proc. Computer Science and Statistic, 13th Symposium on the Interface.

Chambers, J.M., Cleveland, W.S., Kleiner, B. et Tuckey, P.A. (1983): Graphical Methods for Data Analysis, Wadsworth Int. Group, California.

Chernoff, H. et Rizvi, M.H. (1975): Effect on Classification Error of Random Permutations of Features in Representing Multivariate Data by Faces, JASA, 70, 548-554.

Chernoff, H. (1973): Using Faces to Represent Points in k-dimensional Space Graphically, JASA, 68, 361-368.

Cléroux, R. (1982): L'impact présent et futur de l'informatique en statistique, CIPS Review, 18-21.

Cléroux, R., Roy, R. et Fortin, N. (1980): Air Pollution in Montreal: A Statistical Analysis of Sulphur Dioxide Data, Water, Air and Soil Pollution, 13, 143-156.

Cleveland, W.S. et Kleiner, B. (1974): The Analysis of Air Data Pollution From New Jersey and New-York, Annual Meeting ASA, St-Louis, Miss.

Everitt, B.S. (1978): Graphical Techniques for Multivariate Data, Heinemann Education Books, London.

Fienberg, S.E. (1979): Graphical Methods in Statistics, Amer. Statistician 33, 165-178.

Flury, B. et Riedwyl, H. (1981): Graphical Representation of Multivariate Data by Means of Asymmetrical Faces, JASA, 76, 757-765.

Flury, B. (1980): Construction of an Asymmetrical Face to Represent Multivariate Data Graphically, Technical Report No. 3, Université de Berne, Dép. de Statistique.

Friedman, H.P., Farrell, E.J., Goldwyn, R.M., Miller, M. et Siegel, J.H. (1972): A Graphic Way of Describing Changing Patterns, Proc. Comp. Sci. and Statist., 6th. Annual Symposium on the Interface, Berkeley, Calif., 56-59.

Friedman, J.H., et Rafsky, L.C. (1981): Graphics for the Multivariate Two-Sample Problem, JASA, 76, 277-295.

Gascon, A. (1978): Méthodes graphiques d'analyse de données multidimensionnelles, Mémoire de maîtrise, Dép. d'informatique et de recherche opérationnelle, Université de Montréal.

Gnanadesikan, R. (1980): Graphic Data Analysis: Issues, Tools and Examples, Ann. Meeting Amer. Assoc. Adv. Sci., San Francisco.

Gnanadesikan, R. (1977): Methods for Statistical Data Analysis of Multivariate Observations, Wiley.

Hartigan, J.A. (1975): Printer Graphics for Clustering, Journal Statist. Comput. Simul., 4, 187-213.

Hartigan, J.A. (1975): Clustering Algorithms, Wiley.

Jacob, R.J.K. (1980): Correspondence on Fienberg (1979) Amer. Statistician 34, 252-253.

Kalence, K.W. et Kiviat P.J. (1973): Software Unit Profiles and Kiviat Figures, ACM Perf. Eval. Rev. (sept.).

Kent, P. (1982): An Efficient New Way to Represent Multidimensional Data, thèse de doctorat, Ecole Polytechnique Fédérale de Lausanne.

Kleiner, B. et Hartigan, J.A. (1981): Representing Points in Many Dimensions by Trees and Castles, JASA, 76, 260-276.

Kruskal, W. (1977): Visions of Maps and Graphs, Proc. Int. Symp. on Comput. Assisted Cartography, 25-36.

Kruskal, J.B. (1964): Non-Metric Multidimensional Scaling: a Numerical Method, Psychometrika 29, 115-129.

Lee, R.C.T., Slagle, J.R. et Blum, H. (1977): Triangulation Method for the Sequential Mapping of Points from N-Space to Two-Space, IEEE Trans. Comput. 26, 288-292.

Lin, C.H., et Chen, H.F. (1977): Representation-Space Transformation for the Display of Multivariate Chemical Information, Anal. Chem. 49, 1357-1363.

Pickett, R. et White B.W. (1966): Constructing Data Pictures, Proc. 7th. Nat. Symp. on Information Display, 75-81.

Playfair, W. (1801): The Statistical Breviary, London.

Siegel, J.H. Goldwyn, R.M. et Friedman, H.P. (1971A): Iteration and Interaction in Computer Data Bank Analysis: A Case Study in the Physiologic Classification and Assessment of the Critically Ill, Comput. and Biomed. Res., 4, 607-622.

Siegel, J.H., Goldwyn, R.M. et Friedman, H.P. (1971B): Pattern and Process in the Evolution of Human Septic Shock, Surgery, 70, 232-245.

Wainer, H. (1974): The Suspended Rootogram and Other Visual Displays: an Empirical Validation, Amer. Statistician 28, 143-145.

Wainer, H. et Thissen, D. (1981): Graphical Data Analysis, Ann. Rev. Psychol., 32, 191-241.

Wakimoto, K. (1977): Tree Graph Method for Visual Representation of Multidimensional Data, Jour. Japan Statist. Soc., 7, 27-34.

Wakimoto, K. et Taguri, M. (1978): Constellation Graphical Method for Representing Multidimensional Data, Ann. Inst. Stat. Math. 30, 97-104.

Wang, P.C.C. (1978): Graphical Representation of Multivariate Data, Academic Press, N.Y.

Welsch, R.E. (1973): Graphics for Data Analysis, Comput. & Graphics, 2, 31-37.

Liste de publications déjà parues

84-1 DUFRESNOY, A., GAUTHIER, P.M. et OW, W.H. - *Théorème de Runge sur les fermés pour les équations elliptiques*, janvier 1984, 19 pages.

83-24 BENABDALLAH, K. and YOSHIOKA, S. - *On p-large subgroups of p-torsion groups*, décembre 1983, 12 pages.

83-23 DAHEL, Shanoun - *Tests for the mean of a multivariate normal distribution with additionnal information*, décembre 1983, 14 pages.

83-22 DUBUC, S. et TODOR, F. - *La règle du trapèze (3 textes)*, septembre 1983, 29 pa

83-21 ZAIDMAN, Samuel - *Abstract differential equations with almost-periodic solutio* septembre 1983, 13 pages.

83-20 ZAIDMAN, Samule - *Solutions of abstract differential equations with minimal uni form norm*, septembre 1983, 8 pages.

83-19 GAUTRIN, H.-F. et KLEMOLA, T. - *Continuité de la transformation "scattering" po l'équation de Korteweg de Vries*, septembre 1983, 28 pages.

83-18 FRAPPIER, Clément - *Inequalities for entire functions of exponential type*, sept bre 1983, 14 pages.

83-17 DUBUC, Serge et TANGUAY, Monique - *Variables duales dans un programme continu d transport*, août 1983, 13 pages.

83-16 TARDIF, Serge - *Sur la linéarité asymptotique de statistiques de rang signé II*, août 1983, 17 pages.

83-15 FRAPPIER, Clément - *Some inequalities for trigonometric polynomials*, août 1983, 14 pages.

83-14 ARAKELIAN, N.U. ang GAUTHIER, P.M. - *On tangential approximation by holomorphic functions*, juillet 1983, 32 pages.

83-13 BENABDALLAH, K. et BOUANANE, A. - *Sur les tag-modules*, juillet 1983, 14 pages.

83-12 OUDADESS, Mohamed - *Une norme d'algèbre de Banach dans les a.l.u.A-convexes complètes - et - Caractères dans les B_0-algèbres*, juin 1983, 27 pages.

83-11 FRAPPIER, Clément - *On the inequalities of Beinstein-Markoff for an interval*, juin 1983, 20 pages.

83-10 ELVER, E., SARNDAL, C.E., WRETMAN, J.H. and ORNBERG, G. - *Regression analysis ratio analysis for domains, a randomization theory approach*, avril 1983, 25

83-9 FRODA, Sorana - *Comparison of efficient L- and R-estimators of location*, av 1983, 18 pages.

83-8 KILAMBI, Srinivasacharyulu - *On a problem of Kaplansky*, avril 1983, 5 pages.

83-7 DUFRESNOY, Alain - *Sur la stabilité de l'indice d'un point critique non dégén* avril 1983, 12 pages.

93-6 KRAWCEWICZ, Wieslaw - *Sur la méthode de Lyapounov-Schmidt pour les problèmes d bifurcation*, avril 1983, 42 pages.

83-5 DAHEL, S. et GIRI, N.C. - *Quelques distributions dérivant d'une wishart décen* mars 1983, 16 pages.

83-4 BENABDALLAH, K. and YOSHIOKA, S. - *On P-large subgroups of P-torsion reduce groups*, mars 1983, 10 pages.

83-3 DAS, M.N., GIRI, N. and AHMED, M. - *Design through recoding of varietal and level codes*, mars 1983, 17 pages.

83-2 SARNDAL, Carl E. - *Design consistent versus model dependent estimation for sm domains*, janvier 1983, 24 pages.

83-1 DUFRESNOY, Alain - *Un exemple de champ magnétique dans $\mathbb{R}^\nu$* , janvier 1983, 9